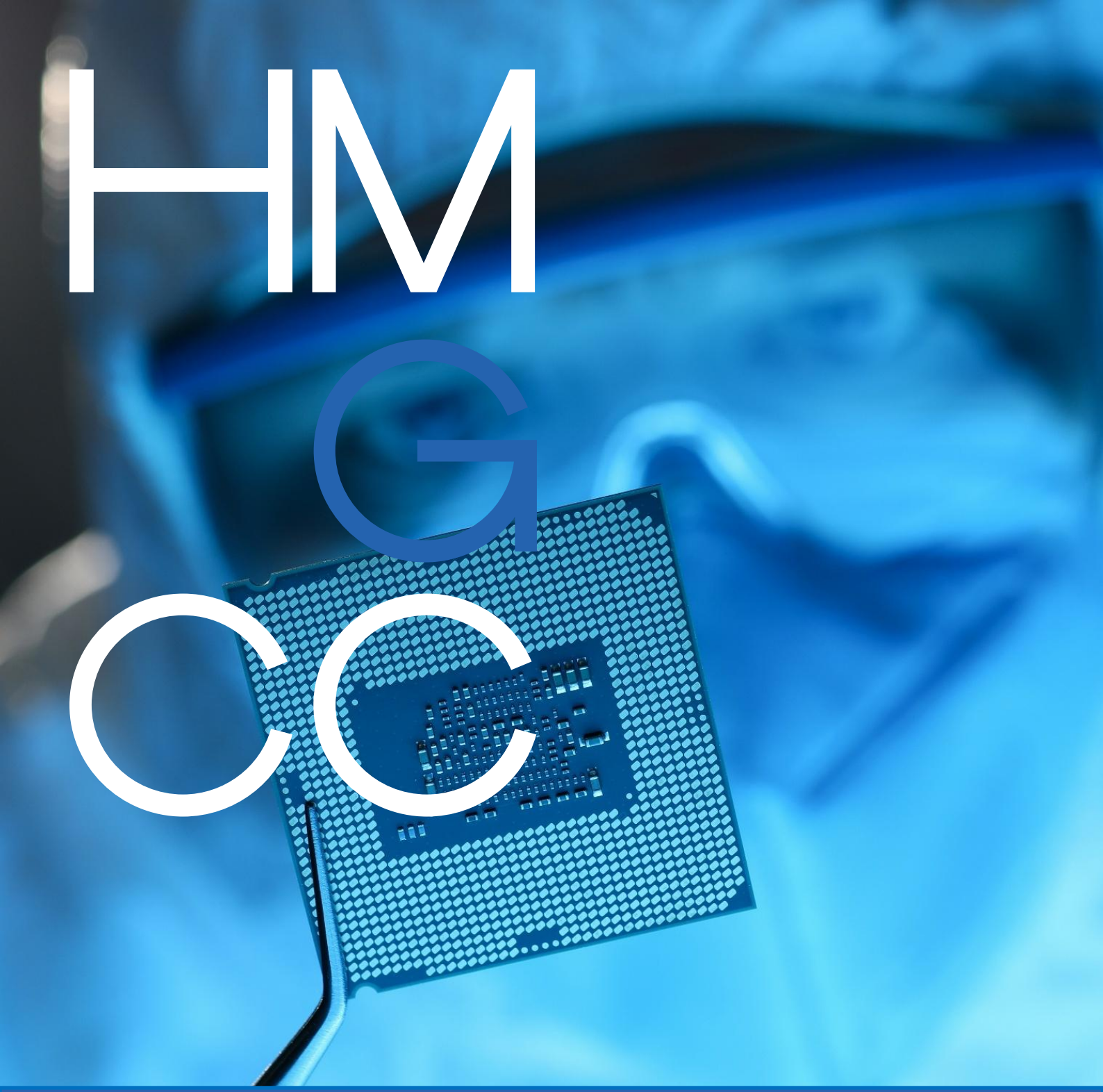

HMGCC Co-Creation

Smart personal assistant for security researchers

Clarification questions and answers

HMGCC
Co-Creation

HMGCC Co-Creation

[dstl]

Document Details: Clarification Q&A in response to the call for proposals

Challenge: Smart personal assistant for security researchers

Deadline for questions: Friday 17th April 2026

#	Question	Answer
1.	Can you please clarify whether collaboration is encouraged, and how you prefer potential collaborators to be included for this competition? Are there any restrictions in terms of their level of participation, eligible project costs, reimbursement percentages and/or ability to co-lead a challenge?	Collaboration is permitted within the specified budget for this challenge. Please ensure that the organisations are in compliance with UK government trade restrictions and/or arms embargoes. Bids comprising one or more organisations must be submitted by a single, accountable lead organisation with whom the Authority would contract (if successful). The teaming and flow-down of terms and payment to other members of the consortia would be the responsibility of the lead organisation. There are no restrictions on the level of participation, eligible project costs or reimbursement percentages.
2.	Are there any hardware specifications for the laptop on which the software tool produced needs to be able to run?	The aim of the laptop is being able to run the solution away from the cloud, and to not depend on a particular closed-source model that could be removed/replaced/tweaked. On this basis, the actual laptop specs don't matter too

		much to us, however an example laptop of 4-8GB VRAM is what may be envisaged. We can self-host larger (e.g. 70B parameter models) with an OpenAI API style endpoint, if model size is critical.
3.	Database & Data Formats: Is there any guidance on the expected structure or format of the test data that will be provided? For instance, whether the document library will consist primarily of PDFs, structured databases, or a mix of formats including images and schematics. Could we see examples of this data?	We expect the test data to comprise a mix of formats, rather than a single structured source. This is likely to include (but not be limited to) documents such as PDFs, Word and Excel files, presentations, and text-based documents, as well as images and diagrams (for example JPEG, PNG, BMP, TIFF), schematics, draw.io outputs, and hand-drawn or scanned material. Recent research projects have not included databases as a searchable source, although it's possible that future projects could involve them. The ability to access a database would be desirable. We don't currently have shareable examples, however the team will provide a small selection of representative documents during the project.
4.	Can you confirm whether the £60,000 maximum budget must include all time, materials, overheads, indirect costs, and any software licensing? The PDF notes funding covers "time, materials, overheads and other indirect expenses," but we want to ensure these are all included within the £60,000 ceiling.	All proposed costs must be included in the maximum £60,000 budget.

HMGCC Co-Creation

[dstl]

5.	Are there required formats or standards for MVP deliverables, code handover, documentation, and confidence-scoring outputs?	There are no mandated standards for the MVPs, however documentation and code should be of a sufficient standard for sharing with HMGCC engineers. This could potentially involve one or two in-person demonstration sessions in Milton Keynes. Where requirement discovery, design and testing is undertaken this should be documented by the Solution Provider.
6.	Do diagrams, tables, and images count towards the six-page limit, and are technical annexes (e.g., architecture diagrams, risk logs) permitted outside that limit?	Diagrams, tables and images count towards the six-page / slide limit. Any information that the bidder would like to be assessed (e.g. technical information or risk logs) should be included within the six-pages / slides. The page/slide limit excludes title pages, references, personnel CVs and organisational profiles.
7.	If the MVP is judged successful, can you outline the likely scope, timelines, and indicative expectations for potential Phase 2 follow-on funding?	There is no commitment for Phase 2 funding. The potential scope, timescales and expectations for follow-on work might be considered based on the outcomes delivered during the current phase.
8.	Are any local servers available for inference compute (running an LLM)? If so, what are the specifications?	We have an OpenAI API endpoint that can host models of at least 70B parameters, and we can also readily host open source models from huggingface.
9.	Is a managed inference platform available for the project (e.g. Government cloud inference API, hosted models, or LLM gateway)? If so, what models are available?	Please see answer to question 8.

HMGCC Co-Creation

[dstl]

10.	What data formats are in scope for analysis or should we assume all; images, pdf, doc, spreadsheet etc.?	We expect the test data to comprise a mix of formats, rather than a single structured source. This is likely to include (but not be limited to) documents such as PDFs, Word and Excel files, presentations, and text-based documents, as well as images and diagrams (for example JPEG, PNG, BMP, TIFF), schematics, draw.io outputs, and hand-drawn or scanned material.
11.	For the scenarios required will sample workflows be provided or should these be researcher-generated based on open-source data?	Sample workflows will be provided during the project.
12.	What audit and explainability requirements apply to autonomous agent decisions - is a human-readable decision log sufficient, or is formal explainability expected?	A succinct human-readable decision log would be fine. Certainly we want to be very clear what's known, what's a guess, and we're interested in solutions that provide clarity here.
13.	Offline Updates: Regarding the desirable requirement to "ensure the software tool remains up-to-date when offline," could you provide further details on the expected mechanism? For instance, does this imply the system should feature a secure, air-gapped update workflow (such as cryptographic integrity verification of update packages provided via USB/SD card), or does this refer to a different functional capability?	The intent of the requirement is twofold: 1) Software/algorithm updates - Users will expect to be able to apply updates, both major and minor, to the assistant itself (runtime, UI, model loading logics etc). This implies a documented, air-gapped update process that validates the integrity and authenticity of any package before it is installed. 2) Data repository updates – the tool should be able to support periodic loading of new offline knowledge bases (e.g. CVE feeds, vendor bulletins etc). This is slightly

		vague as we don't always know what data repositories will be available offline.
14.	Code Analysis Expectations: Could you please provide additional context on the level of code analysis the system is expected to perform? It would be helpful to know the typical characteristics of the target code (e.g., expected codebase size, single script vs. repo) and the specific types of security analysis findings or architectural insights the assistant must be capable of generating from it.	With the recent fast paced change going on in the LLM world, code analysis isn't a priority for this solution. At most, it would be nice to get some basic architectural insights, such as • High level component design that lists • Firmware modules (bootloader, kernel, drivers, OTA updater) • Communication interfaces (UART, i2c, SPI, ethernet, Wi-FI etc) • Trusted and untrusted execution domains • External dependencies (libs, 3rd party SDKs) • Identifies Attack surfaces (Ports, exposed APIs, file-system mounts)

15.	**Question:** Cultural Bias Mitigation: Could you please elaborate on the desirable requirement to "recognise and mitigate cultural biases to ensure a nuanced understanding"? Specifically, at what stage of the workflow is this most relevant, and what does mitigation look like in practice? For example, should the LLM chat interface proactively correct a user's implicit assumptions, or should the system provide feedback during data ingestion (e.g., advising the user to seek out Dutch sources if a machine's manufacturer is Dutch but only English documents were provided)? A concrete example of this requirement in action would be highly appreciated. **HMGCC answer:** As a starting point we encourage teams to review the Responsible Technology Adoption Unit (RTAU) blog, which contains useful background material on Bias Mitigation: https://rtau.blog.gov.uk/category/bias/ A few notes that may also help: Taking the Dutch manufacturer example, the system could add an alert or tag against the corpus upload that recognises the Dutch origin but the English only documentation. A simple flag to the user. Ability to run a corpus level audit on upload producing a summary, including any bias information. Addition of a check list to follow up; 'add Dutch security advisories', 'include Chinese CERT notices' For retrieval and reasoning, a smaller prompt focus report or coverage label flagging 'English only'. Ability to automatically check previous uploads for alternative-cultural sources. Interaction/UI – it would be useful to be checked on language, again Dutch example; 'ThermoCo firmware weakness – only English documentation has been loaded, consider obtaining Dutch security advisories'

Within this type of feature, allowing the users to decide how pro-active the tool is might be nice.

Pro-active: Assistant automatically adds extra source (or prompt the user to get it) and corrects biased statements without asking

Suggestive: Prompts to alert the user that the tool has identified potential biases

OFF: Vanilla tool – we don't worry about bias as we know the sources are limited currently.

Example

Pattern	Action	User sees
Ingestion	User previously uploaded a folder ThermoCo_T200/ that contains only ThermoCo_T200_EN.pdf. Ingestion engine tags it lang=EN, origin=NL. It notes 'Missing Dutch source -> 0/2 Dutch documents available in local repo'	Modal banner at the top of the chat window. "All uploaded documents for ThermCo are English. Two Dutch advisories exist locally (See Cultural-Gap tab) (if not local source, just the message about 100% English for a Dutch manufacturer)
Retrieval	Query vector-search returns 1. English NVD entry (CVE-2023-1234) 2. English vendor advisory (ThermoCo-ReleaseNotes) 3. (no Dutch docs) Diversity score = 100% English	System builds the answer and appends coverage label: Coverage: English 90%/Dutch 0%

HMGCC Co-Creation

[dstl]

	Interaction Policy = PROACTIVE	System looks for missing Dutch documentation in its known sources/repos. It finds ThermoCo_T200_NL_Bulletin_2023_02.pdf. Offline MT engine translates the relevant paragraph into English and adds it to the response, providing a provenance link/alert	The answer information, including the translated paragraph, with the provenance information
16.	I am a final year MSc Computer Science student at [ ] University. My thesis supervisor has just shared your Smart personal assistant for security researchers challenge with me and I was struck by the challenge description, as the technical requirements mirror the exact problem set I am tackling for my dissertation. I know you are looking for a solution provider to build this to TRL 6, but I'd love to help the winning team once they are picked in May/June. If there is any way you could pass my details to them as a potential unpaid programmer or collaborator, I'd be incredibly grateful. I am familiar with your IP terms - I just would love to contribute to the project and see this technology actually work in the real world to support national security.	Thank you for your interest. In the spirit of collaboration, we are happy to make the successful bidder aware of your offer, but please note that this is without expectation or commitment. Any resulting collaboration agreement would be the responsibility of yourself and the successful Solution Provider.	

HMGCC Co-Creation

[dstl]

17.	I note that the budget is £60K. Are you expecting the output of this project to hit TRL 6 at the end of the 12 weeks? If so, do you not think that this is an unrealistic amount, especially considering that the ideas may be at significantly lower TRLs? £60K will not allow us to put on a proper team to get this to high TRL. It might get us to TRL 4 at best especially considering that this is a tough problem to crack.	Proposals will be assessed on desirability, feasibility and viability in accordance with the evaluation criteria set out in the Challenge Form.
18.	I have a pdf version of the challenge statement which appears to be an image of the document, not viable text and links.	The Challenge Form is available on the HMGCC website: Challenges | HMGCC Challenge Portal
19.	Technical scope: What types of industrial control systems should the test data represent? (e.g., PLCs, SCADA, DCS, additive manufacturing controllers, or a mix?)	All of them or any of them. The ICS is an example domain, any domain with hardware/software components are in scope.
20.	Technical scope: Can you provide more detail on the expected volume and variety of test data? (e.g., approximate number of documents, file types, total data size)	We expect the test data to comprise a mix of formats, rather than a single structured source. This is likely to include (but not be limited to) documents such as PDFs, Word and Excel files, presentations, and text-based documents, as well as images and diagrams (for example JPEG, PNG, BMP, TIFF), schematics, draw.io outputs, and hand-drawn or scanned material.

21.	Technical scope: Will the test data include actual binary firmware or compiled code, or only source code and documentation about code?	HMGCC test data will not include binary firmware or source code.
22.	Technical scope: For 'handwritten annotations' — should the tool handle standalone handwritten notes (e.g., scanned notebook pages), or annotations overlaid on existing documents (e.g., handwriting on a printed schematic)?	Ideally both, but we'd take annotations on overlaid on an existing document over standalone handwritten note.
23.	Offline and hardware: What are the target laptop specifications? (e.g., minimum RAM, GPU availability, OS — Windows/Linux/macOS, storage constraints)	The aim of the laptop is being able to run the solution away from the cloud, and to not depend on a particular closed-source model that could be removed/replaced/tweaked. On this basis, the actual laptop specs don't matter too much to us, however an example laptop of 4-8GB VRAM is what may be envisaged. We can self-host larger (e.g. 70B parameter models) with an OpenAI API style endpoint, if model size is critical.
24.	Offline and hardware: Is there a preferred or mandated operating system for the deployment environment?	Windows 11 or Ubuntu (latest).
25.	Offline and hardware: Are there any security constraints on the software that can be installed on	There are no security constraints.

	the target laptop? (e.g., restrictions on Docker, Python runtimes, specific libraries)	
26.	Evaluation and testing: How will the tool be evaluated during in-house testing at HMGCC? Are there specific test scenarios, benchmarks, or acceptance criteria beyond the stated requirements?	We will use it on actual mission problems similar to the use case described and evaluate qualitatively whether it saved time.
27.	Evaluation and testing: For the confidence scoring requirement — is there a preferred scoring methodology or threshold framework, or is this for the solution provider to define?	Happy for the provider to define their scoring/threshold framework.
28.	Evaluation and testing: What does 'validate responses before publishing' mean in practice? Is this an automated self-check, a human-in-the-loop review step, or both?	Similar to the answer for Question 16. Useful to be able to have a choice – automated self-check, or listing sources for a human review check.
29.	Delivery and process: What is the expected format and cadence for sprint reviews during the 12-week project? (e.g., bi-weekly demos, written reports, or both?)	We are open to bidders proposing an appropriate cadence for Sprint reviews, however the majority of our 12-week projects are typically 3 x 4-week sprints.
30.	Delivery and process: Will HMGCC provide access to subject matter experts (security researchers) for user testing and feedback during the project?	Yes.

HMGCC Co-Creation

[dstl]

31.	Delivery and process: Is the test data available at project kick-off, or will it be provided incrementally?	We expect the test data to comprise a mix of formats, rather than a single structured source. This is likely to include (but not be limited to) documents such as PDFs, Word and Excel files, presentations, and text-based documents, as well as images and diagrams (for example JPEG, PNG, BMP, TIFF), schematics, draw.io outputs, and hand-drawn or scanned material. Recent research projects have not included databases as a searchable source, although it's possible that future projects could involve them. The ability to access a database would be desirable. We don't currently have shareable examples, however the team will provide a small selection of representative documents during the project.
32.	IP and integration: Are there any preferred or restricted open-source licences for components used in the solution? (e.g., restrictions on GPL, AGPL)	No, we have no preferred (or restrictions on) open-source licences.
33.	IP and integration: Is there an expectation that the tool will eventually integrate with existing HMGCC systems or workflows, and if so, are there any interface standards to be aware of?	Not in scope.

34.	Beyond RAG: When you say proposals should go beyond off-the-shelf RAG — could you elaborate on specific limitations of current RAG approaches that you have experienced or are concerned about?	The poor handling of structured artifacts, difficulty representing hierarchies such as SBOM graphs, firmware memory maps etc. Lack of multimodal ingestion – project data is in many file formats. Lack of Bias control or awareness – currently upload only documents in English, the tool assumption is English is the only relevant language (See Question 16) Updates can be a brittle, error prone process.
35.	Beyond RAG: Is there interest in the tool building structured representations (e.g., a component knowledge graph or system architecture map) from the ingested data, or is the focus primarily on conversational Q&A over unstructured sources?	Both.
36.	What is the target laptop specification for the offline analysis tool, particularly available RAM and whether a GPU (Apple Silicon unified memory, or NVIDIA GPU) is available?	The aim of the laptop is being able to run the solution away from the cloud, and to not depend on a particular closed-source model that could be removed/replaced/tweaked. On this basis, the actual laptop specs don't matter too much to us, however an example laptop of 4-8GB VRAM is what may be envisaged.

		We can self-host larger (e.g. 70B parameter models) with an OpenAI API style endpoint, if model size is critical.
37.	Is there a preferred or required operating system for the delivered tool (Linux, Windows, macOS)?	Windows or Ubuntu.
38.	When will test data be provided to successful applicants? Early access at or before project kick-off would significantly help to de-risk the 12-week delivery.	We don't currently have shareable examples; however, the team will provide a small selection of representative documents during the project. This will be within the first 4 weeks.
39.	What formats and scale should we expect in the test data -- approximately how many documents, what mix of PDFs, images, schematics, code, and forum posts, and how many distinct components are in the target system?	We expect the test data to comprise a mix of formats, rather than a single structured source. This is likely to include (but not be limited to) documents such as PDFs, Word and Excel files, presentations, and text-based documents, as well as images and diagrams (for example JPEG, PNG, BMP, TIFF), schematics, draw.io outputs, and hand-drawn or scanned material. Recent research projects have not included databases as a searchable source, although it's possible that future projects could involve them. The ability to access a database would be desirable.

40.	Can the test data be referenced or used during the pitch presentation, or is it only available post-award?	We don't currently have shareable examples, however the team will provide a small selection of representative documents during the project. This will be within the first 4 weeks.
41.	One potential architecture could include an online ingestion phase where open-source (publicly available, non-sensitive) source material is processed to produce an enriched corpus, which is then transferred to an air-gapped laptop where all analysis occurs offline. Using this approach, the system could benefit from the use of frontier AI models on non-classified data (which are currently significantly more capable than the most capable open source models). Is this approach acceptable, or is HMGCC expecting a fully offline pipeline from ingestion onwards?	Yes, potentially acceptable, although the benefit is not immediately clear from the question.
42.	Are there constraints on which open-source AI models can be used for the offline analysis component? For example, are models of Chinese provenance (DeepSeek, Qwen) acceptable for a Phase 1 prototype, given that the models are open weights, and the tool operates on open-source data?	There are no constraints on the models used. Documentation covering what models are in use will be required.
43.	If usage of online frontier AI models is permitted (in contrast to offline open weights models), are there constraints on which commercial AI models can be	It is not permitted.

	used, which data centre locations can be used, and a ceiling on cost for these models? I.e. -- you could get better results if you used Max / xhigh thinking, but these models cost substantially more (e.g. $30 input / $180 output+ for GPT 5.4 Pro)	
44.	We intend to submit an application through [X], a company incorporated in the United Kingdom. For transparency, the founder of the company holds dual [Y and Z] nationality and has been resident in the UK for more than five years. Could you please confirm whether, on that basis, [X] would be eligible to apply for this challenge?	At an organisational level, please ensure that the organisations are in compliance with UK government trade restrictions and/or arms embargoes. We may however, request the project team undertake BPSS checks or equivalent.
45.	The challenge specifies that the software tool must operate without an internet connection. Could you confirm whether the system is expected to run entirely on a single, user-managed laptop (e.g., with local LLM inference and indexing), or is it acceptable for the tool to be deployed across multiple devices within a single, air-gapped network (e.g., a laptop connected to a dedicated, offline compute server) where the laptop acts only as a secure client via a physically isolated, encrypted connection?	The aim of the laptop is being able to run the solution away from the cloud, and to not depend on a particular closed-source model that could be removed/replaced/tweaked. On this basis, the actual laptop specs don't matter too much to us, however an example laptop of 4-8GB VRAM is what may be envisaged. We can self-host larger (e.g. 70B parameter models) with an OpenAI API style endpoint, if model size is critical.

		We can also readily host open source models from huggingface.
46.	How much access to HMGCC security researchers and SMEs can solution providers expect during the 12-week project? I expect significant benefit will be gained from co-design with domain experts.	As a minimum, HMGCC SME input will be provided during the Sprint Planning and Review ceremonies. In practice, we also recognise the value of collaborative working and will endeavour to provide additional SME input and steers during the Sprints where practicable.
47.	Will sprint reviews involve HMGCC end-users, or only the Co-Creation delivery team?	HMGCC SME input will be provided during the Sprint Planning and Review ceremonies.
48.	Could you clarify the expected operating environment in terms of hardware constraints (e.g. typical RAM/VRAM ranges), and whether you anticipate the system adapting its capabilities dynamically based on available resources?	The aim of the laptop is being able to run the solution away from the cloud, and to not depend on a particular closed-source model that could be removed/replaced/tweaked. On this basis, the actual laptop specs don't matter too much to us, however an example laptop of 4-8GB VRAM is what may be envisaged. We can self-host larger (e.g. 70B parameter models) with an OpenAI API style endpoint, if model size is critical. We can also readily host open source models from huggingface.

		The laptop would be used by one person at a time. We don't have a set of standard tests to evaluate the solution you provide. Our Security Researchers will evaluate the MVP by testing it in real-life against a representative problem.
49.	In practice, will analysts always operate in a fully self-contained environment, or are there scenarios where access to a secure local or on-premises compute resource is expected?	The Security Researchers will have access to secure, on-premises compute.
50.	Are we correct in assuming the primary user is an experienced security researcher, or should the system also accommodate less specialised users such as operators or engineers?	The only users of the solution will be trained security researchers.
51.	From your experience, which stages of the research process are most time-intensive — locating relevant material, interpreting complex artefacts such as schematics, or validating conclusions? Would it be possible to share representative examples of the types of artefacts analysts work with (e.g. manuals, schematics, teardown images), or at least describe their typical structure and variability?	Project and technology dependant. Most recently the interpretation of the data has taken more time.

52.	In your experience, do the most critical insights tend to reside in well-structured documentation (e.g. datasheets), or in less structured and more ambiguous sources such as diagrams, scans, or forum discussions?	Project dependant. We have recent projects where the technical pdfs and subject matter specifications have been the most helpful. In another project, it was a series of forum posts that prompted a rethink of our findings and a redirection in focus. And the right diagram is nearly always helpful.
53.	When the system encounters ambiguity or incomplete evidence, would you prefer it to present a single, qualified interpretation, or to surface multiple plausible hypotheses with supporting evidence?	We would prefer the tool to surface multiple plausible hypotheses with supporting evidence.
54.	How would you expect the system to behave in cases where it cannot confidently resolve a query for example, should it explicitly defer, suggest further avenues of investigation, or highlight gaps in available data?	It should explicitly defer, suggest further avenues of investigation, or highlight gaps in available data.
55.	At the conclusion of the 12-week phase, what would constitute a meaningful improvement in researcher effectiveness from your perspective?	This will be a qualitative assessment based on using the proposed solution at that point.
56.	Are there particular shortcomings or failure modes in existing tools that you would consider important to avoid in this context?	The failure to recognise ‘No information’, or ‘Not enough information’ to provide a good answer. Prompting the user if they want to continue.

		Ability to interrupt a 'thinking' moment when the tool is clearly going the wrong way, or the prompt has been lacking.
57.	Are there any constraints or expectations around how updates or improvements to the system would be delivered within an offline or restricted environment?	If it is easier for you to deliver the solution physically on a laptop, please do so. Alternatively, we can load your software by using vSphere. This applies to the initial installation and also for updates.
58.	What are the target hardware specifications for the host laptop, such as minimum RAM, GPU/VRAM requirements, and storage type, and is there a preferred operating system for the final TRL 6 deliverable?	The actual laptop specs don't matter too much to us, however an example laptop of 4-8GB VRAM is what may be envisaged. We can self-host larger (e.g. 70B parameter models) with an OpenAI API style endpoint, if model size is critical. We can also readily host open source models from huggingface. Windows or Ubuntu are the preferred operating systems. The laptop would be used by one person at a time.
59.	Does "without an internet connection" imply a completely standalone device with no network interface, or can the tool reside on an internal air-gapped LAN? What is the approved process for	If it is easier for you to deliver the solution physically on a laptop, please do so. Alternatively, we can load your software by

	initially side-loading the software and model weights onto these offline devices?	using vSphere. This applies to the initial installation and also for updates.
60.	Regarding "Technical Credibility" and the "Chain of Trust," does HMGCC prioritise compiled, self-contained solutions with minimal external runtime dependencies and a locked-down, auditable dependency tree over interpreted architectures with larger attack surfaces?	For the prototype, we prioritise a well-documented and auditable dependency tree over a fully compiled architecture.
61.	To address the "vast quantities of data" requirement, is it acceptable for the solution to support a hybrid data model? Specifically, can the tool ingest and query pre-processed data libraries optimised on high-performance infrastructure, while also maintaining the capability to ingest and index new, local files on-the-fly as a researcher discovers them during a tear-down?	Yes, a hybrid solution would be acceptable.
62.	Regarding the desirable requirement for "periodic updates" and long-term viability, will preference be given to architectures that allow for rapid, low-cost updates of the knowledge base through re-indexing, rather than solutions that require time-intensive and computationally expensive model retraining or fine-tuning to incorporate new technical specifications?	A solution that works effectively once installed and can be updated quickly and cheaply through re-indexing would be a preferred approach. We recognise that model retraining or fine tuning is a significant undertaking in an air-gapped environment and would want to understand how any proposal manages this practically and cost effectively. Innovative approaches that can be demonstrated are welcome.

63.	To ensure the "memory of queries" remains intact over several weeks, does HMGCC prefer architectures that utilise embedded, crash-resilient storage, such as ACID-compliant local databases, to prevent data corruption during the frequent power cycles common in field-based research?	Yes, it would be preferable for the solution for maintain conversation history and query memory in a way that survives power cycles and unexpected shutdowns. ACID compliant databases would be a sensible, understood approach but, again, those decisions are up to the provider.
64.	Since the tool will be used by researchers on a laptop alongside other complex tasks, will the evaluation favour solutions that maintain a low background memory and CPU footprint to prevent resource contention, or is the assessment indifferent to the hardware overhead of the assistant?	For practical purposes with this challenge, we've focused deliberately on 'laptop' solutions to encourage self-contained solution that we will be able use immediately. As mentioned on the briefing call, we do have an air-gapped infrastructure. A solution architected in a manner to allow it to scale to a more powerful deployment environment would be viewed positively. To this end the resource contention on the laptop is not something we'd evaluate (unless the tool crashes it constantly). If the solution is designed to scale up, then we'd need to consider its hardware/resource requirements.
65.	Requirement 87 mentions characterisation from multimedia inputs like schematics and handwritten annotations. Should the natural language function be capable of visual grounding, specifically referencing and reasoning over spatial elements	Text extraction and indexing from images and schematics is a baseline essential. We would view positively a solution which demonstrates spatial awareness of diagrams; recognising that components are

	and annotations within images, or is the requirement focused primarily on text-based extraction?	connected, or hand-written annotations points to a specific element. Full visual grounding and reasoning over complex engineering schematics would be amazing. We are realistic about the challenge this represents within the project constraints, and welcome proposals that are honest about what they can do and with what level of confidence.
66.	Are there specific restrictions regarding the use of open-weights foundational models as a base, and will preference be given to model-agnostic architectures that allow for the swapping of model weights to integrate newer or more specialised models without requiring a rewrite of the core application logic?	There are no restrictions on the use of open weights foundational models as a base. A model agnostic architecture offers flexibility and longevity, and we would view this positively.
67.	While the current requirement specifies a laptop, should the architecture consider future portability to mobile platforms like tablets or iPads, and are there specific security constraints for such mobile hardware in this context?	The actual laptop specs don't matter too much to us, however an example laptop of 4-8GB VRAM is what may be envisaged. Mobile hardware must not have a connection to the outside world (this includes Bluetooth) and cannot have a camera or microphone.
68.	Is this call open for a consortium of industry/academia?	Consortium bids are allowed within the total £60k limit.

69.	Is there a need for export functionality form obsidian, one note etc?	We would view positively any solution that can import from popular note taking and knowledge management tools, such as one note or Obsidian etc, as researchers may have existing notes they wish to incorporate. However, whilst this is a useful capability, it this isn't an explicit requirement.
70.	Corporate databases — format clarification The essential requirements include "corporate databases" as an input type. Could you clarify what formats or access patterns are in scope? For example: structured database exports (CSV, SQL dumps, XML), direct database connections, or proprietary enterprise system exports?	Recent research projects have not included databases as a searchable source, although it's possible that future projects could involve them. The ability to access a database would be desirable. We are not restricted to any particular database format and to not have a preference re connection or ingestion methods.
71.	Test data availability and timing The brief states that test data will be provided. Will the test data be available at project kick-off, or should teams plan for a lead time? Is a representative sample or format specification available before the project starts to allow teams to prepare ingestion pipelines?	We don't currently have shareable examples; however, the team will provide a small selection of representative documents during the project. This will be within the first 4 weeks.
72.	Initial setup environment The system must operate offline in production. Will there be a connected environment available for	The deliverable must be fully self-contained (i.e. with no online connection) from first use.

	initial setup (e.g. downloading model weights, container images, and dependencies), or must the deliverable be fully self-contained on physical media from first use?	
73.	TRL 6 acceptance criteria Could you clarify the specific TRL 6 acceptance criteria for this challenge? We interpret TRL 6 as "system/subsystem model or prototype demonstration in a relevant environment" per the UKRI definition. Is the relevant environment a researcher's laptop running against provided test data, or are there additional environment constraints?	Yes – this would be a researcher's laptop running against provided test data, in an air-gapped environment.
74.	Will you provide GPU infrastructure to deploy the solution? So infrastructure is out of the budget?	During the 12-week project, the bidder would need to be able to demonstrate the solution working on their own test environment that is sufficient to operate the prototype at TRL 6. The eventual target operating environment (i.e. after the conclusion of this project, and at higher levels of TRL7+) would run on HMGCC hosted infrastructure.
75.	What can you tell us about the volume and type of the test data provided for this challenge? How closely does it match the documentation in the field?	We expect the test data to comprise a mix of formats, rather than a single structured source. This is likely to include (but not be limited to) documents such as PDFs, Word and Excel files, presentations, and

		text-based documents, as well as images and diagrams (for example JPEG, PNG, BMP, TIFF), schematics, draw.io outputs, and hand-drawn or scanned material. Recent research projects have not included databases as a searchable source, although it's possible that future projects could involve them. The ability to access a database would be desirable. We don't currently have shareable examples, however the team will provide a small selection of representative documents during the project.
76.	The challenge document notes that you would like to see proposals that do not focus on off-the-shelf RAG. Could you say more about what you are looking for? Are there specific limitations of standard RAG approaches that you have encountered, or particular capabilities you consider essential that they lack?	The poor handling of structured artifacts, difficulty representing hierarchies such as SBOM graphs, firmware memory maps etc. Lack of multimodal ingestion – project data is in many file formats. Lack of Bias control or awareness – currently upload only documents in English, the tool assumption is English is the only relevant language (See Question 16). Updates can be a brittle, error prone process

HMGCC Co-Creation

[dstl]

77.	Would there be any opportunity during the 12-week project to engage briefly with an operational security researcher for feedback on the tool? Even a short review session would substantially strengthen the user-centred design of the final prototype.	As a minimum, HMGCC SME input will be provided during the Sprint Planning and Review ceremonies. In practice, we also recognise the value of collaborative working, and will endeavour to provide additional SME input and steers during the Sprints where practicable.
78.	My understanding is that TRL 6 implies that the solution needs to be more of a proof-of-concept than a "read-to-use" capability. What are your priorities in terms of the scope? I.e., are you primarily looking to validate that the product design fits into the user process, or that the quality of retrieval and information ingestion / structuring is production-grade, or that the latencies are viable, or anything else? As an example, would you prefer a solution that focuses on the quality retrieval and delivers minimal UX matching the bid requirements, or a solution that focuses on delivering an optimal UX while using off-the-shelf indexing and retrieval?	We are looking for a prototype that works well, for example, by ingesting various types of data and providing the ability for sensible interactive queries. We would prefer a small, well rounded and performant MVP as a deliverable, as opposed to a larger well-polished solution that isn't as performant under-the-hood.
79.	I want to know if an individual researcher working as a research fellow in Europe (outside the UK) is eligible to participate in this competition.	At an organisational level, please ensure that the organisations are in compliance with UK government trade restrictions and/or arms embargoes. We may however, request the project team undertake BPSS checks or equivalent.

80.	Are there minimum performance expectations in output speed? (e.g., latency or tokens per second)?	We don't have a set of standard tests to evaluate the solution you provide. Our Security Researchers will evaluate the MVP by testing it in real-life against a representative problem.
81.	Are there constraints on model origin (e.g., open-weight only, region-specific restrictions)?	No.
82.	Do any AI models need to run fully offline on a laptop?	Yes.
83.	Retrieval-Augmented Generation (RAG) will likely be needed as memory is assumed to be tight. You want more than standard RAG as detailed in the brief. What are your expectations around answer accuracy and hallucination prevention?	The minimum expectation is that every answer is grounded in cited sources from the loaded materials. Uncertainty is prominently flagged and the tool acknowledges when it does not have sufficient information. For this challenge we are pragmatic about hallucination prevention and answer accuracy. Our evaluation will be qualitative and a tool that says 'Not enough information' is more valuable than a tool that gives a confident wrong answer.
84.	Assuming a solution where responses will be strictly grounded in retrieved source data only this tends towards a higher assurance level than standard graph RAG approaches, is the	Grounding responses in retrieved source data as a design philosophy would be preferred. A combination of rule-based checks and AI assisted judgement is

HMGCC Co-Creation

[dstl]

	expectation deterministic verification mechanisms (e.g., schema-constrained outputs, rule-based checks, or enforced source citation)?	acceptable, provided the researcher can see what is known, what is inferred and what is uncertain.
85.	What volume of data should the system handle for file inputs (approximate size, #sources)?	Data is likely to be uploaded in smaller groupings as the researchers gather information sources. At minimum I would expect to be able to upload 50 large pdf and/or word docs, excel workbooks, several hundred photographs, schematics, and image types.
86.	Can you provide sample user prompts, questions?	Please see question 15.
87.	Are existing Question–Answer (QA) pairs likely to be available?	No
88.	Are there example reports where QA pairs could be generated?	No
89.	Point 4 "cross checking" seems to contradict the need to run without internet connection.	The intent of the cross-checking requirement is verification within the loaded corpus rather than live external queries.

HMGCC Co-Creation

[dstl]

90.	Do you already have defined success criteria for the challenge in terms of measuring response accuracy/recall/precision/hallucination rates etc?	We don't have a set of standard tests to evaluate the solution you provide. Our Security Researchers will evaluate the MVP by testing it in real-life against a representative problem.
91.	How will Phase 1 deliverables be evaluated? Gold-standard query set supplied, built collaboratively, or left to supplier?	We don't have a set of standard tests to evaluate the solution you provide. Our Security Researchers will evaluate the MVP by testing it in real-life against a representative problem.
92.	Can the solution be integrated into an existing agent platform that we already have, with the new tech for the solution being an added component?	Yes. Please note the system will need to be designed to be air-gapped.
93.	Is pulling information from publicly available sources (either automatically, or human-directed) part of the required specification, or is it assumed that the researcher will find and upload the required manuals, specs etc?	The researcher be responsible for finding and uploading the relevant materials.
94.	Is there cost auditing for this contract in the same way as for Innovate grants / contracts, or do we have full discretion how to allocate costs?	Proposals should be fully costed and submitted on a fixed, firm price basis. The selected supplier will be required to invoice a fixed amount per sprint, with payment milestones to be agreed during the contracting stage.

HMGCC Co-Creation

[dstl]

95.	Part 4. of essential requirements - is it a requirement that the solution pull the industry publications / research etc?	No, as above, the researcher will be responsible for finding and uploading the relevant materials.
96.	Do you evaluate the team and / experience and existing clients and capabilities of the applicants? If so, where is the best place to include this in the application?	Diagrams, tables and images count towards the six-page / slide limit. Any information that the bidder would like to be assessed (e.g. technical information or risk logs) should be included within the six-pages / slides. The page/slide limit excludes title pages, references, personnel CVs and organisational profiles.
97.	Could it be possible to know the hardware available to run the app?	The aim of the laptop is being able to run the solution away from the cloud, and to not depend on a particular closed-source model that could be removed/replaced/tweaked. On this basis, the actual laptop specs don't matter too much to us, however an example laptop of 4-8GB VRAM is what may be envisaged. We can self-host larger (e.g. 70B parameter models) with an OpenAI API style endpoint, if model size is critical.

HMGCC Co-Creation

[dstl]

		We can also readily host open source models from huggingface.
98.	[If the user needs to] check internet forums in a different computer? How does the person upload the information in this system?	Data from forums could be manually inputted into the system by the Security Researcher or captured in screen shot and then ingested into the smart personal assistant.
99.	Will the data be provided before we apply?	We don't currently have shareable examples; however, the team will provide a small selection of representative documents during the project. This will be within the first 4 weeks.
100.	Would you consider a solution that utilised a MacOS based laptop?	No.
101.	Will the slides also be made available after the briefing?	No.
102.	Is there list of equipment you typically procure?	No. Any of the main manufacturers of CNC machinery, or control equipment could be considered. HMGCC maintains a standard manufacturing setup including CNC Mills, Lathes, Grinding machines, Laser etc. We must stress Industrial Control Systems is used as an illustrative example domain

		within this challenge. The domain encompasses a broad range of technologies including but not limited to programmable controllers, human machine interfaces, supervisory control and data acquisition systems and a variety of machinery types such as precision manufacturing equipment, additive manufacturing systems and process control machinery. These systems typically involve a mix of physical hardware components, embedded processors, communication interfaces and software layers. They communicate using a range of industrial protocols and expose a variety of physical and logical interfaces which may represent potential attack surfaces. Any of the main manufacturers of CNC machinery, or control equipment can be considered. HMGCC maintains a standard manufacturing setup and CNC Mills, Lathes, Grinding machines, Laser etc. This challenge is not limited to the ICS domain. Any complex system with hardware and software components, documentation across multiple formats and a mix of open and proprietary interfaces is in scope.

HMGCC Co-Creation

[dstl]

103.	Can you share the list of vendors and type of equipment's generally procured.	No. Any of the main manufacturers of CNC machinery, or control equipment could be considered. HMGCC maintains a standard manufacturing setup including CNC Mills, Lathes, Grinding machines, Laser etc. We must stress Industrial Control Systems is used as an illustrative example domain within this challenge. The domain encompasses a broad range of technologies including but not limited to programmable controllers, human machine interfaces, supervisory control and data acquisition systems and a variety of machinery types such as precision manufacturing equipment, additive manufacturing systems and process control machinery. These systems typically involve a mix of physical hardware components, embedded processors, communication interfaces and software layers. They communicate using a range of industrial protocols and expose a variety of physical and logical interfaces which may represent potential attack surfaces. Any of the main manufacturers of CNC machinery, or control equipment can be considered. HMGCC maintains a standard

		manufacturing setup and CNC Mills, Lathes, Grinding machines, Laser etc. This challenge is not limited to the ICS domain. Any complex system with hardware and software components, documentation across multiple formats and a mix of open and proprietary interfaces is in scope.
104.	The brief mentions that one of the data sources is forums. Does that mean the tool needs to be able to retrieve data from online sources, or will the information from online sources be provided to the tool in some other way, since the brief also states the tool is expected to work offline?	Data from forums could be manually inputted into the system by the Security Researcher or captured in screen shot and then ingested into the smart personal assistant.
105.	Will ground truth data be supplied as part of the test data so that we can evaluate what we build?	We don't have a set of standard tests to evaluate the solution you provide. Our Security Researchers will evaluate the MVP by testing it in real-life against a representative problem.
106.	Scale of data: Approximately how large is the test dataset in terms of number of documents and total file size?	We don't currently have shareable examples; however, the team will provide a small selection of representative documents during the project. This will be within the first 4 weeks.

HMGCC Co-Creation

[dstl]

107.	Confidence scoring: Is there a preferred methodology or benchmark for the confidence scoring mechanism, or is this left to the solution provider?	This is at the discretion of the Solution Provider.
	Multimodal inputs: Does image understanding need to handle technical schematics specifically (e.g. circuit diagrams), or general photographs of teardowns?	Both are relevant to our use cases, and we would expect the tool to handle both to the best extent possible within the project constraints. Technical schematics, we'd expect a minimum accurate text and label extraction. Proven spatial understanding of connections and component relationships would be viewed as a strong differentiator. For teardown photographs we would expect the tool to identify and describe visible components, interfaces and physical features accurately. We understand the different level of technical challenge around these areas and welcome proposals that are honest about what they can achieve for each input and at what confidence level.
108.	Are there any tools or platforms already in use at HMGCC that the solution must integrate with or avoid?	Integration is not in scope.

109.	Will any data or outputs generated during the 12-week project remain solely with you, or can applicants use anonymised technical artefacts (e.g., extraction pipeline outputs, not source documents) to improve the harness/tooling post-project?	The data and outputs generated by the HMGCC researchers will remain solely with us. We will provide a write up sharing our findings based on our testing and evaluation process.
110.	Would you accept anything less than TRL6? It's a lot of scope for limited budget.	We are aiming for TRL 6. Proposals will be assessed on desirability, feasibility and viability in accordance with the evaluation criteria set out in the Challenge Form.
111.	Internet Access During a Research Session It is very clearly stated in the brief that the tool must operate without an internet connection. What is less clear is whether the researcher's working environment itself is air-gapped (i.e. they have no internet access at all during an assessment), or whether they can continue to browse on trusted online forums and add newly found documents to the tool as the investigation progresses. If the researcher can indeed do this, what is the preferred method of new information ingestion to the air-gapped device?	The solution itself will need to operate as if it is air gapped. The Security Researcher can browse on trusted online forums via different methods and can add newly found documents to the tool as the investigation progresses. Data from forums could be manually inputted into the system by the Security Researcher or captured in screen shot and then ingested into the smart personal assistant.
112.	Test Data Could you give any indication of the approximate format and volume of the test data that will be provided for the task — for example, the number of	We expect the test data to comprise a mix of formats, rather than a single structured source. This is likely to include (but not be limited to) documents such as PDFs, Word and Excel files, presentations, and

	documents, predominant file types (PDF, images, code languages, videos, audio files etc.) and whether it covers a single product or multiple?	text-based documents, as well as images and diagrams (for example JPEG, PNG, BMP, TIFF), schematics, draw.io outputs, and hand-drawn or scanned material. Recent research projects have not included databases as a searchable source, although it's possible that future projects could involve them. The ability to access a database would be desirable. We don't currently have shareable examples, however the team will provide a small selection of representative documents during the project.
113.	Target Hardware Specification Does HMGCC have a target device specification for the offline deployment environment? For instance, is there a preferred operating system, GPU provider (e.g. will the device be CUDA-enabled), or is there a specific budget allocated to the device? The brief mentions a laptop – would the HMGCC be open to alternative hardware solutions?	The aim of the laptop is being able to run the solution away from the cloud, and to not depend on a particular closed-source model that could be removed/replaced/tweaked. On this basis, the actual laptop specs don't matter too much to us, however an example laptop of 4-8GB VRAM is what may be envisaged. We can self-host larger (e.g. 70B parameter models) with an OpenAI API style endpoint, if model size is critical. We can also readily host open source models from huggingface.

HMGCC Co-Creation

[dstl]

114.	Translation Scope The desirable requirements mention translating and indexing non-English data sources. Are there particular languages that are most relevant? This would help us scope which offline translation models to use, and therefore, properly assess whether this would be a risk to the downstream accuracy of the system.	Useful languages to start with may be: German, Japanese, French, Chinese.