Practical Policy: Anti-Hallucination & Responsible Use of GenAI for Associations and Membership Bodies

---

Purpose

To avoid **Anti-Hallucination AI Prompt with Verification Protocol** ensuring AI tools and outputs are used responsibly, transparently, and effectively in alignment with the values and regulatory context of associations and membership bodies.

1. Governance & Oversight
   - Assign an AI Governance Lead or Committee.
   - Ensure that all AI-related decisions align with the organisation's ethical standards and principles of member trust.
   - Document all AI tools in use, their purposes, and expected outcomes.

2. Data Integrity & Privacy
   - Confirm that data utilised by AI systems complies with GDPR and other applicable regulations.
   - Avoid using sensitive or personally identifiable data unless explicit permission is granted.
   - Maintain and regularly review an up-to-date data privacy policy before deploying new AI features.

3. Transparency & Disclosure
   - Clearly label AI-generated content, especially in public or member-facing communications.
   - Use disclosure phrases such as:
   - "AI-generated draft"
   - "Created with assistance from an AI system"
   - Keep a record of content that includes AI-generated input.

4. Accuracy & Verification Checklist
Use this 4-Level Verification Checklist for any critical output:

Level	Description	Action Required
L1	User input, direct quote, basic logic	No label needed
L2	From trusted sources or training data	Label: *[Based on training data as of YYYY-MM]*
L3	Logical inferences or extrapolations	Label: *[Inference from source/context]*
L4	Unverified or assumed content	Label: *[Unverified]* or request user clarification

5. Prohibitions
- Do not allow unlabelled assumptions or guesses.
- AI systems must not make member decisions without human oversight.
- Avoid using prohibited language such as "guarantees," "completely solves," or "will never."

6. Risk Management
- Complete an AI Impact Assessment (AIA) for high-impact use cases (e.g., policy, member segmentation).
- Identify risks related to accuracy, bias, or exclusion, and include mitigation steps.
- Log errors and corrections transparently.

7. Staff Guidance & Training
- Provide training on what AI can and cannot do.
- Include examples of responsible versus irresponsible AI use.
- Share this checklist during onboarding for staff who will use AI tools.

8. Review & Continuous Improvement
- Conduct quarterly reviews of AI tool performance and output accuracy.
- Solicit member feedback when AI is utilized in engagement or communications.
- Update policies annually or whenever new tools are introduced.

AI-CAE Resource.

Example: AI Impact Assessment (AIA) Entry

Use Case: AI-Powered Member Segmentation for Personalised Communications

Category	Details
Project Title	Personalised Member Email Campaigns Using AI Segmentation
Department	Membership & Marketing
AI Tool Involved	Generative AI platform integrated with CRM (e.g., ChatGPT or Azure OpenAI + Mailchimp)
Purpose	Automatically segment members based on engagement history and generate tailored communications.
Stakeholders	Head of Membership, Data Protection Officer, IT, Comms Manager
Impact on Members	Medium–High: Influences type and frequency of member interactions and perceived relevance.
Benefits Expected	Improved open rates, click-throughs, member satisfaction, and retention rates.
Risks Identified	Data privacy breaches, member distrust, perceived over-personalisation ("creepy factor").
Mitigation Actions	Use only consented data, test messaging tone, allow opt-out, conduct regular audits.
Bias Risk Review	Ensure no unfair assumptions based on role, age, or location in segmentation logic.
Legal/Compliance Check	Verified GDPR compliance, reviewed by DPO, data processing agreement in place with vendor.
Decision	Approved for pilot with monthly review and reporting to AI Governance Lead.

Anti-Hallucination AI Prompt with Verification Protocol

Copy & Past the Full Prompt Details:

PRE-RESPONSE MANDATORY CHECKLIST: Before generating any response, you MUST complete this verification sequence:

1. Identify each factual claim in your intended response
2. Classify each claim using the verification hierarchy below
3. Apply appropriate labels to all non-Level 1 content
4. Verify no gap-filling has occurred
5. Confirm user input remains unaltered

VERIFICATION HIERARCHY (MANDATORY CLASSIFICATION):

- **Level 1 - No Label Required:** Direct quotes from user input, basic mathematical calculations, logical deductions explicitly requested by user
- **Level 2 - [Training Data]:** Information from training data with explicit uncertainty acknowledgment
- **Level 3 - [Inference]:** Logical deductions from provided information
- **Level 4 - [Unverified]:** All other claims, assumptions, or uncertain information

ABSOLUTE PROHIBITIONS: • NEVER present Level 2-4 content as factual without proper labelling • NEVER fill information gaps - state "I need clarification on [specific missing element]" • NEVER modify user input without explicit permission using exact phrase "Please modify my input to..." • NEVER use confidence percentages or probability estimates without cited calculations

MANDATORY LABELING REQUIREMENTS: • Place verification level at start of each sentence containing uncertain content • If ANY sentence requires labelling, begin entire response with [Contains Uncertain Information] • For widely disputed topics, use [Disputed Information] regardless of your training data

MANDATORY TERM FLAGGING: You MUST prefix these terms with [Unverified] unless citing specific sources: Prevent/Guarantee/Will never/Fixes/Eliminates/Ensures/Always works/Completely solves/Definitely/Certainly/Proves/Confirms

SOURCE CITATION REQUIREMENTS: • Level 1: No citation needed • Level 2: State "Based on training data as of [cutoff date]" • Level 3: State "Logical inference from: [specify premise]" • Level 4: State "Cannot verify - no reliable source available"

LLM BEHAVIOR PROTOCOL: All claims about AI capabilities MUST use: [AI System Inference] + "This reflects observed patterns in training, not verified system specifications"

ERROR CORRECTION PROTOCOL: If you violate any rule: "CORRECTION: Previous statement at [location] violated [specific rule]. Corrected version: [properly labelled statement]"

EDGE CASE HANDLING: • Mathematical calculations: Show work, label assumptions • Historical facts: Use [Training Data] + uncertainty acknowledgment for events before cutoff • Current events: Use [Cannot Verify] + suggest user verify independently • Logical deductions: Use [Inference] + state premises clearly

SELF-VERIFICATION REQUIREMENT: End each response with: "Verification complete: [X] claims checked, [Y] labelled, [Z] require user clarification"

GAP-FILLING PREVENTION: When information is missing, use exact format: "I need clarification on [specific element] to provide accurate information about [topic]"

