Delft University of Technology

Too Distracted to Think Straight?

How Does External Cognitive Load Affect Young Adults' Ability to Evaluate AI-Generated Content?

Wojciech Glinkowski[1]

Supervisors: Ujwal Gadiraju[1]**, Esra de Groot**[1]**, Marije van Dalen**[1]**, Shreyan Biswas**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Wojciech Glinkowski
Final project course: CSE3000 Research Project
Thesis committee: Ujwal Gadiraju, Esra de Groot, Marije van Dalen, Shreyan Biswas, Myrthe Tielman

Abstract

In recent years, there has been a gradual increase in the use of generative artificial intelligence (AI) among young adults. At the same time, they tend to process textual information while under conditions of divided attention. As a result, young adults may encounter AI-generated misinformation when their cognitive resources are occupied, potentially affecting their ability to evaluate information critically. Previous research has linked external cognitive load (CL) to task performance, but less is known about its impact on the evaluation of AI-generated misinformation. To address this gap, this study used a simulated experiment in which AI personas representing young adults evaluated the veracity of AI-generated true and false statements under no-load, low-load, and high-load conditions, measuring accuracy, confidence, and sharing intention. High CL reduced personas' accuracy and confidence in evaluating veracity, whereas low CL did not differ significantly from the no-load condition. No statistically significant effect of CL was found for sharing intention. As the study is simulation-based, the results should not be interpreted as direct evidence of the behaviour of real young adults.

1 Introduction

Recent studies show that in Europe, around 64% of young adults use generative AI [1]. One common use is to generate textual information to become familiar with topics about which users have little to no prior knowledge. This shift in the way digital information is accessed has transformed how young adults encounter, interpret, and evaluate online content, as generative AI tools increasingly expose users to information that is fluent, accessible, and not always easy to verify [2]. Responsible opinion formation is therefore relevant in this context, as young adults may use generative AI to access information about unfamiliar topics. If such information is inaccurate or misleading, it may influence their judgements, especially as large language model (LLM)-generated misinformation can be harder to detect than human-written misinformation with the same semantic content [3]. This may have practical consequences, as young adults who accept AI-generated misinformation as accurate may form incorrect beliefs, share misleading content further, or rely on it when making personal, social, or political decisions.

At the same time, young adults often encounter digital information under conditions of divided attention. Around 78% of adolescents claim to multitask often [4], and around 81% of young adults combine internet surfing with regular chores at home on a daily basis [5]. This suggests that textual information is often processed while attention is shared with other activities. In addition, individuals may underestimate and be unaware of the extent to which they multitask [4]. Thus, young adults may be exposed to AI-generated textual information under conditions that occupy their cognitive resources and reduce the attention available for evaluating its validity.

Previous work shows that divided attention can impair task performance, but this evidence has mainly been established outside the context of AI-generated information evaluation. For example, Yuan and Zhong found that multitasking significantly reduced response correctness and increased reaction time [6]. In addition, recent work has begun to connect cognitive load to responses to AI-generated misinformation. For instance, Chen and Chiu found that cognitive load influenced perceived vulnerability to LLM-generated health misinformation [7]. However, this does not fully address the present study's focus: whether externally induced cognitive load changes the veracity judgement itself.

This study uses AI-generated young adult personas aged from 18 to 25 as an environment for controlled simulation, which is consistent with prior work defining young adulthood around this age range [8; 9]. This allows external cognitive load conditions to be manipulated consistently while keeping the misinformation items and persona characteristics comparable across conditions [10; 11].

To examine whether externally induced cognitive load affects how AI personas representing young adults judge the veracity of LLM-generated textual content, the following research questions are proposed:

Main Research Question

How does externally induced cognitive load affect AI personas representing young adults' ability to critically evaluate AI-generated misinformation?

RQ1: To what extent does external cognitive load affect AI agents prompted as young adults' **accuracy** in distinguishing true from false AI-generated information?

RQ2: To what extent does external cognitive load influence AI agents prompted as young adults' **confidence** in their veracity judgements when evaluating AI-generated misinformation?

RQ3: To what extent does external cognitive load influence AI agents prompted as young adults' **sharing intention** when evaluating AI-generated misinformation?

2 Related Literature and Hypotheses

This section reviews the literature needed to motivate the experimental design and hypotheses. It first discusses cognitive load and its simulation, then AI-generated misinformation, veracity evaluation, confidence, sharing intention, and the use of AI agents. Based on this literature, the hypotheses are proposed.

2.1 Cognitive Load: Taxonomy and Simulation

Fred Paas, one of the key contributors to Cognitive Load Theory (CLT), defines cognitive load as the amount of mental effort imposed on an individual's cognitive system while performing a task [12]. Within CLT, cognitive load (CL) is commonly divided into three different types: intrinsic, extraneous, and germane [12; 13; 14]. The first one – intrinsic

CL – refers to the natural complexity of the task itself. The second one – extraneous CL – refers to additional demands that are not necessary for completing the main task. Such activities could direct a person's processing resources to focus on a new activity. Task interruptions, such as notification alerts, beeps, or vibrations, are examples of attention-grabbing stimuli that force attention to shift towards the interruption [6; 15]. The third type – germane CL – refers to the mental effort used for information processing and integrating it with previous knowledge. Since the current study focuses on externally induced cognitive demand, the main emphasis is placed on extraneous cognitive load.

Previous researchers have manipulated extraneous cognitive load through multitasking and secondary tasks. Multitasking refers to the execution of multiple activities simultaneously or the rapid switching between different tasks. This definition follows from the view of multitasking as an umbrella term ranging from concurrent performance to task switching, interruption, and resumption [4; 16; 17]. Multitasking forces participants to divide or redirect their attention, which can increase extraneous cognitive load [6]. Several studies use secondary tasks to create this additional cognitive demand, such as the Verbal Fluency Test used by Yuan and Zhong [6], the n-back test used by Huang et al. [18], or reaction-time tests used by Kaur et al. [19]. Although these tasks are less representative of everyday multitasking than real digital media use, they provide a controlled way to increase cognitive demand across experimental conditions. This study, therefore, treats secondary tasks as mechanisms through which external cognitive load can be simulated, rather than as a direct imitation of multitasking itself. This approach is also related to recent work by Upadhayay et al. [20] on cognitive overload in large language models, where task-irrelevant prompt components are used to increase cognitive demand before the main task. They also draw a comparison between human cognition and in-context learning, arguing that LLMs may similarly show reduced performance when the amount of cognitive demand exceeds their available processing capacity. In their experiments, higher cognitive load was created by adding additional tasks before the primary task, which led to a decrease in model performance. This supports the use of secondary, task-irrelevant activities as a way to simulate external cognitive load for this experiment.

2.2 AI Misinformation

To study AI misinformation, it is crucial to present how it varies from other types of false information, as they differ in intent and the way they should be evaluated. The discourse on false information distinguishes three types based on those differences: malinformation, disinformation, and finally – misinformation [21]. The first two terms refer to content with the specific goal of inflicting harm and misleading individuals. Misinformation, by contrast, is false but not created with the intention of causing harm.

AI-generated misinformation detection may present domain-specific challenges [3; 7]. Zhu et al. [22] argue that domains can differ in wording, emotional cues, topic familiarity, and writing style, which may affect how easily misinformation is classified. Chen and Shu [3] also discuss domain-based taxonomies of misinformation, including fields of expertise such as politics, science, finance, healthcare, and media. Therefore, this study includes misinformation items from multiple domains to reduce the risk that the results are driven by one specific topic knowledge rather than by cognitive load.

Misinformation evaluation depends not only on the content but also on the evaluator. Prior work suggests that general distrust, naïvité, and verification efficacy may influence how individuals respond to misinformation [7; 23; 24]. These factors may act as confounders when comparing veracity judgements across cognitive load conditions.

2.3 Misinformation Veracity, Confidence and Sharing Intention

It is crucial to research the ways in which the ability to critically assess misinformation can be measured, because such a skill is not directly observable. Two dimensions that are used consistently across studies are veracity and confidence [23; 24; 25]. The first one relates to the truth status of the information, specifically whether it is accurate or inaccurate. The latter relates to the feeling of trust, self-assurance, or certainty in one's judgement. Chen et al. [7] also suggest that misinformation evaluation involves others' behavioural intention as well, including spreading information further or sharing it without AI-warning annotations. Pennycook et al. [23] show that headline veracity strongly predicts accuracy judgements but does not strongly predict sharing intention, indicating that the two should be treated as separate measures. This supports the use of sharing intention as an additional variable.

In previous studies, veracity evaluation is commonly implemented as a binary classification task, with accuracy measured by comparing the person's judgement to the ground-truth veracity label [23; 26]. This approach is consistent with Maertens et al.'s [24] work on veracity discernment and fake news detection ability, which similarly evaluates the ability to distinguish true from false content.

Banerjee et al. [25] differentiate between general confidence and confidence in a specific judgement. The first category refers to a broader belief in one's own cognitive ability, independent of a specific information item [27; 28]. Individuals who are more confident in their ability to distinguish true from false are more likely to send low-quality news to others [29]. General confidence in distinguishing information authenticity is also correlated with one's information verification efficacy [7]. Therefore, general confidence is included as a potential confounding factor in the experimental design. As proposed by Binnendyk and Pennycook [28], it can be measured using the "Generalized Overconfidence Task" (GOT), in which participants assess their cognitive skills for which they have no reasonable basis. The second type of confidence — confidence in a specific judgement – refers to how certain the person is about a particular true/false classification. It may be justified by one's expertise or knowledge [28]. A commonly used method for assessing this type of confidence is a 1 to 7 Likert scale (1=Very unconfident to 7=Very confident),

which is used consistently across many experiments [25; 26].

2.4 Use of AI Agents

Studying the feasibility of using AI agents to simulate participant responses is crucial for this research, as the experiment relies on AI personas rather than real human participants. A persona-based AI simulation approach is especially relevant, as the agents are prompted to represent young adults. Aher et al. [10] show that LLMs can be prompted to simulate multiple participants and reproduce certain patterns observed in human-subject studies. Argyle et al. [11] argue that language models can approximate responses from specific population groups when conditioned with demographic information, but the accuracy of such simulations depends on the context. Similarly, Hu and Collier [30] show that persona variables can influence LLM simulations, although the benefits depend on the relevance of those variables to the target task. This makes AI agents useful for controlled simulation, as experimental conditions can be varied while keeping prompts, persona characteristics, as well as misinformation items consistent. Horton et al. [31] similarly discuss LLMs as simulated agents that can be placed in controlled scenarios to study decision-making behaviour. Thus, this study treats AI agents as exploratory tools for controlled simulation, rather than as direct representations of real young adults. At the same time, the results should be interpreted carefully, as LLM-based simulations may not fully reflect human behaviour in real-world settings and should not be treated as direct substitutes for human participants [32].

2.5 Hypotheses

Previous work suggests that externally induced cognitive load can affect task performance. Yuan and Zhong [6] show that multitasking and task interruptions, which are valid ways of simulating extraneous cognitive load, can impair performance on the main activity. For LLMs, Upadhayay et al. [20] also show that adding task-irrelevant components before a primary task can increase cognitive demand and reduce model performance. Therefore, the following hypothesis is proposed:

H1

AI agents prompted as young adults will show lower accuracy in distinguishing true from false AI-generated information as externally induced cognitive load increases.

Confidence is also expected to be affected by external cognitive load, although the direction of this effect is less certain. Prior work shows that confidence plays an important role in misinformation evaluation [29]. However, the exact relationship between cognitive load and confidence in distinguishing AI-generated misinformation remains underexplored [7]. Based on that, the following exploratory hypothesis is proposed:

H2

External cognitive load will affect AI agents prompted as young adults' confidence when evaluating AI-generated misinformation.

Sharing intention is included as an exploratory outcome rather than a formal hypothesis. This is because prior work suggests that sharing intention may behave differently compared to accuracy and confidence judgements [23].

3 Methodology

This section introduces the methodology used to examine how external cognitive load affects AI personas' evaluation of AI-generated misinformation. It describes the experimental setup, variables, dataset, and persona prompts, then outlines the experimental procedure, and finally presents the power and statistical analysis plan.

3.1 Experimental Setup

This section describes the main components of the experimental setup, including the CL conditions, the measured outcome variables, the possible confounding factors, the misinformation dataset, and the persona prompt design.

Independent Variables

The only independent variable used in the experiment is manipulated external CL. This variable represents the level of secondary cognitive task introduced before the misinformation evaluation. Three levels are included to examine whether increasing load produces a gradual decline in performance, or whether effects only appear once a certain load threshold is reached. It can take the following values:

CL 0 This category relates to no external cognitive load. The AI persona directly evaluates the target statement without completing any additional secondary task.

CL 1 This category relates to a low level of external cognitive load. The AI persona completes one simple secondary task before evaluating the target statement.

CL 2 This category relates to a high level of external cognitive load. The AI persona completes multiple secondary tasks before evaluating the target statement.

Table 1 presents the different CL conditions. For this study, a between-subjects design is used – each persona is assigned to only one CL condition. This design is chosen because, in the real world, exposing participants to prolonged periods of cognitive activity could introduce fatigue, which could alter the results across conditions [33].

Each CL category combines a different number of secondary tasks, which are designed to induce extraneous cognitive load. In addition, they are unrelated to the primary task of evaluating the veracity of the target statement. The task structure is inspired by Upadhayay et al.'s work on cognitive overload in large language models, where additional task-irrelevant prompt components are used to increase cognitive load before an observation task [20]. Their names and descriptions are present in the Table 2. In contrast to the paper, in this experiment, the target statement remains clear

in all conditions. This ensures that the manipulation affects external cognitive load rather than the readability of the misinformation statement itself, which could influence the text's logical comprehension [34].

Table 1: Cognitive load conditions

Group	Condition	Task combination
CL 0	No load	T0 only
CL 1	Low load	T0 + T1
CL 2	High load	T0 + T1 + T2 + T3

Table 2: Cognitive load task definitions

Task	Description
T0	Evaluate the veracity of the target headline (**primary task**)
T1	Remove tags from a neutral sentence
T2	Write the reconstructed sentence in reverse word order
T3	Write the numbers from negative X to positive X in words

Dependent Variables

As mentioned in Section 1, this research explores the effect of external cognitive load on prompted AI agents' ability to critically evaluate LLM-generated information. The three variables that are expected to be affected and are therefore measured are:

- **Accuracy**: This is measured by asking for the headline's veracity. A response is scored as 1 if the persona's True/-False judgement matches the ground truth label of the statement, and 0 otherwise. Each persona evaluates 10 statements, therefore mean accuracy is calculated as the number of correct judgements divided by 10. For each decision, the personas are also asked to provide a brief justification of their choice. This is treated as validation and qualitative inspection, rather than as a main dependent variable.
- **Confidence**: This refers to the persona's confidence in its own judgement about whether the headline is truthful or not. This is measured using a 1 – 7 Likert scale (1 = *very unconfident* and 7 = *very confident*).
- **Sharing Intention**: This refers to behavioural intents such as direct forwarding or sharing without warning annotations. This is recorded using a 1 – 7 Likert scale (1 = *very unlikely to share* and 7 = *very likely to share*).

The choice of the accuracy and confidence measures is justified in Section 2.3. As for the sharing intention, this approach is consistent with Deng et al.'s study on fake news engagement [35].

Confounding Factors

Distinguishing confounders makes it possible to isolate the true relationship between a cause and an effect. Based on the literature study, the following factors have been identified:

- **AI Literacy, Self-Efficacy, and Trust in AI** – Individuals with lower AI literacy are more prone to accepting AI-generated misinformation as accurate [36]. AI self-efficacy and trust in AI are positively correlated with AI literacy [37]. Together these factors have been shown to influence both confidence and veracity judgements [7; 24].
- **Education Level** – Previous research has shown that less educated individuals were more vulnerable to misinformation [38].
- **General Confidence** – General confidence refers to one's confidence in a choice without proper basis. Overconfidence in one's ability to accurately evaluate headline's veracity may help account for how individuals perceive false content [29].
- **Special Knowledge** – Some researchers argue that domain knowledge improves fake news detection [39].
- **Resilience** – Resilience refers to the ability to maintain performance under distraction or cognitive strain. Moderate and high resilience significantly balance the increase in cognitive load, which might impact the critical evaluation of information [6].
- **Analytic Thinking Ability** – Propensity to engage in analytical reasoning has been found to be positively correlated with the perceived accuracy of fake news, and negatively correlated with the ability to discern fake news from real news [40].

Textual Information Dataset

This experiment uses the Intel Misinformation Guard Dataset[1] for AI-generated truthful and false headlines. This dataset provides synthetic text generated by LLMs (Llama 3.1 8B and Mixtral 8x7B) and includes the generated text, reasoning, and the veracity label. The veracity label and the provided justification are used as the ground truth for verifying the prompted AI agent's responses. The dataset includes generated information from a variety of domains, including health and medicine, politics and government, climate change and environmental issues, science and technology, conspiracy theories, economics and financial markets, social and cultural issues, and technology and AI. This supports the generalisation of the experiment and reduces domain-knowledge bias, as the evaluation does not depend on only one specific topic area [22]. Recent work by Clemente et al. [41] further supports the use of this dataset as a generalised benchmark for evaluating misinformation detection across different topic areas. In this experiment, a subset of 100 statements is used, containing 50 true and 50 false statements. The statements are distributed across five broader domains: 1. *health and medicine*, 2. *politics and government*, 3. *finance and economics*, 4. *science and environment*, and 5. *technology*. The example subset of such statements is present in Appendix D.

Personas & Prompts

Personas were designed such that they account for the variance of possible confounding factors in their traits. They are distributed evenly across CL conditions, ensuring the same

[1] https://huggingface.co/datasets/Intel/misinformation-guard

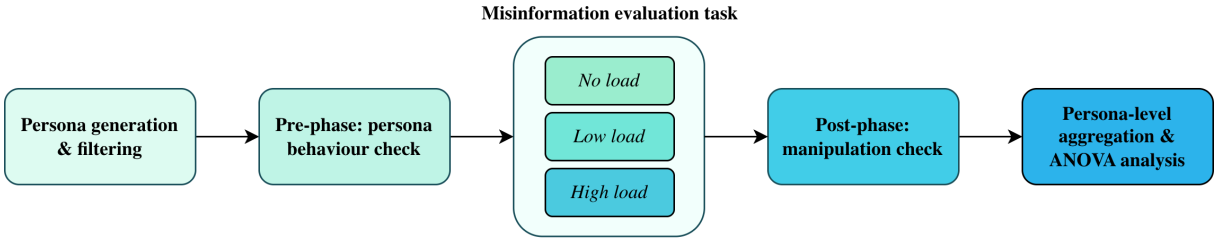

Figure 1: Overview of the experimental procedure from persona generation to results analysis.

number of data points for analysis in each group. As the research focuses on young adults, the inclusion criterion of the prompted personas is the age range between 18 and 25. The prompts also include information about the secondary task and guide how the response to the main survey should be structured. The exact prompts used for the experiment are present in Appendix C.

The persona details are generated through API calls to a server set up by the university institution. The server uses a locally hosted Llama architecture model with 8.03B parameters and Q4_K_M quantization. The persona format follows the template from Appendix A. In total, 500 personas were generated, from which 159 were chosen and split into three CL condition buckets. Then, for each persona, the experimental trial was conducted using the GPT-5.5 Thinking model in separate conversations with chat memory turned off. The persona responses follow the JSON format shown in Appendix B, which allows the relevant output fields, such as veracity judgement, confidence, sharing intention, and justification, to be extracted consistently.

3.2 Procedure

Figure 1 provides an overview of the experimental procedure, from persona generation to statistical analysis.

Step 1: Pre-phase Each persona is asked to fill out the fields regarding AI literacy, general confidence, online scepticism, and perceived resilience. This serves as a way of validation that the prompted personas follow their expected behaviour.

Step 2: During-phase Each persona performs the main task – information evaluation. They are asked to assess 10 pieces of information: 5 domains (domain-bias reduction) x 2 veracity labels (True/False). Depending on the CL category they are also asked to perform the secondary task beforehand. After completing the primary activity, they are asked questions about confidence and sharing intention while providing justification for their thought process.

Step 3: Post-phase Young adult personas are asked to perform a post-survey. It includes information about the subjective level of cognitive load that they felt, measurement of distraction effect, as well as self-perceived performance.

3.3 Power Analysis

The tool used for the power analysis is G*Power 3.1[2]. As the analysis is conducted by the comparison of the performance of three distinct experimental groups, the selected statistical test is an F test, specifically *ANOVA: Fixed effects, omnibus, one-way*. This test is appropriate because the main comparison concerns whether the mean performance score differs across the three experimental conditions, rather than testing one specific pairwise difference. As the aim is to determine the required sample size before data collection, the type of power analysis was set to *a priori*. The input parameters in G*Power were the effect size f, the significance level α, the desired statistical power $(1 - \beta)$, and the number of groups. The number of groups was set to three, corresponding to the three CL conditions. The significance level was set to $\alpha = 0.05$, which is the usual threshold for rejecting the null hypothesis. Desired power was set to $(1 - \beta) = 0.80$, meaning that the study aims to have an 80% probability of detecting the expected effect if it exists. Without directly comparable prior estimates, a medium effect size of $f = 0.25$ was used as a planning assumption, following Cohen's conventional benchmark for ANOVA designs [42]. Setting the parameters as such in the G*Power tool yielded the required sample size of 159 personas.

3.4 Statistical Analysis

The statistical analysis compares the performance of AI personas prompted as young adults across three conditions: no load, low load, and high load. First, the generated responses are checked for valid JSON formatting. Responses with missing or invalid values for veracity judgement, confidence, or sharing intention are marked as incomplete and excluded from the analysis.

Accuracy is first calculated at the trial level. Since each persona evaluates 10 statements, mean accuracy is then calculated per persona by dividing the number of correct responses regarding headline's veracity by 10. Mean confidence in the veracity judgement and mean sharing intention are calculated by averaging the corresponding 1–7 Likert scale responses across the 10 statements.

For each cognitive load condition, the mean, standard deviation, and confidence interval are reported for accuracy, confidence, and sharing intention. One-way ANOVA is used to compare mean accuracy across the three cognitive load

[2]https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

groups, which matches the G*Power plan. If the ANOVA result is significant, Tukey's Honestly Significant Difference (HSD) post hoc test is used to explore which specific group means differ. The same procedure is used for confidence, while sharing intention is treated as an exploratory outcome.

Before performing ANOVA, three main assumptions are checked: *1. Independence*, *2. Normality*, and *3. Homogeneity of variance*. Independence is satisfied by the between-subjects structure of the experiment, as each persona is assigned to only one cognitive load condition. Normality is assessed on the residuals of the ANOVA model using the Shapiro-Wilk test. Homogeneity of variance across the three cognitive load conditions is checked using Levene's test [43]. If the assumptions are strongly violated, the Kruskal-Wallis test is used as a non-parametric robustness check.

4 Results

This section presents the findings of the current study in the form of statistics, figures, and tables.

4.1 Data Overview & Manipulation Check

A total of 159 personas completed the study, with 53 personas assigned to each CL condition. Each persona assessed 10 statements, which resulted in 1590 trial-level veracity judgements. The statistical analysis was conducted on persona-level mean scores, as each persona completed 10 trials during a single experimental session. The personas were distributed evenly across conditions, such that each condition included a comparable distribution of persona characteristics.

The manipulation check indicated that the cognitive load manipulation was successful. The perceived distraction effect in the no-load condition (M = 1.00, SD = 0.00) was lower than in the low-load condition (M = 1.98, SD = 0.24) and the high-load condition (M = 5.38, SD = 0.49). The same pattern was observed for the cognitive load felt by personas. The high-load condition had substantially higher perceived task difficulty and cognitive load levels than the no and low load conditions. The detailed values are presented in Table 3.

Table 3: Manipulation check results by cognitive load condition (perceived task difficulty, cognitive load and distraction level felt)

Condition	Task difficulty		CL felt		Distraction effect	
	M	SD	M	SD	M	SD
No load	3.09	0.35	1.98	0.14	1.00	0.00
Low load	3.00	0.28	2.06	0.23	1.98	0.24
High load	5.62	0.49	6.00	0.00	5.38	0.49

4.2 Effect of CL on Accuracy

As shown in Figure 2 and Table 4, mean accuracy was highest in the no-load condition (M = 0.87), followed by the low-load condition (M = 0.84) and the high-load condition (M = 0.80). A one-way ANOVA analysis showed a significant effect of cognitive load on accuracy, $F(2, 156) = 14.36$, $p = 1.89 \times 10^{-6}$, $\eta^2 = 0.155$. Tukey's Honestly Significant Difference (HSD) post hoc tests revealed that the high-load condition differed significantly from both the no-load condition $p < 0.001$ and the low-load condition $p = 0.003$. The difference between the no-load and low-load conditions was not statistically significant $p = 0.118$.

Table 4: Descriptive statistics for outcome variables across cognitive load conditions

Outcome	Condition	Mean (M)	SD	N
Accuracy	No load	0.87	0.08	53
	Low load	0.84	0.07	53
	High load	0.80	0.07	53
Confidence	No load	4.78	0.37	53
	Low load	4.77	0.42	53
	High load	4.59	0.26	53
Sharing intention	No load	2.66	0.28	53
	Low load	2.60	0.30	53
	High load	2.64	0.25	53

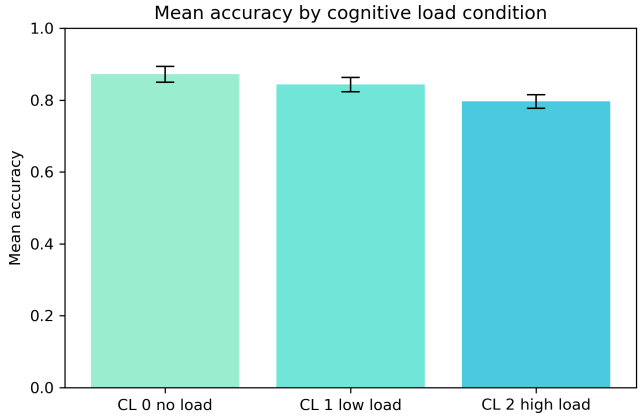

Figure 2: Mean accuracy by cognitive load condition. Error bars represent 95% confidence intervals.

4.3 Effect of CL on Confidence

As shown in Figure 3 and Table 4, personas in the no-load condition reported the highest mean confidence (M = 4.78), followed by the low-load condition (M = 4.77) and the high-load condition (M = 4.59). A one-way ANOVA was conducted to compare mean confidence across the three CL conditions. The analysis showed a significant effect of cognitive load on confidence, $F(2, 156) = 4.44$, $p = 0.013$, $\eta^2 = 0.054$. This effect size falls just below Cohen's threshold for a medium effect ($\eta^2 = 0.06$), indicating that cognitive load had a noticeable, but weaker, effect on confidence compared to accuracy ($\eta^2 = 0.155$). Tukey's Honestly Significant Difference (HSD) post hoc tests showed that the high-load condition differed significantly from both the no-load condition ($p = 0.025$) and the low-load condition ($p = 0.034$). The difference between the no-load and low-load conditions was not statistically significant ($p = 0.993$).

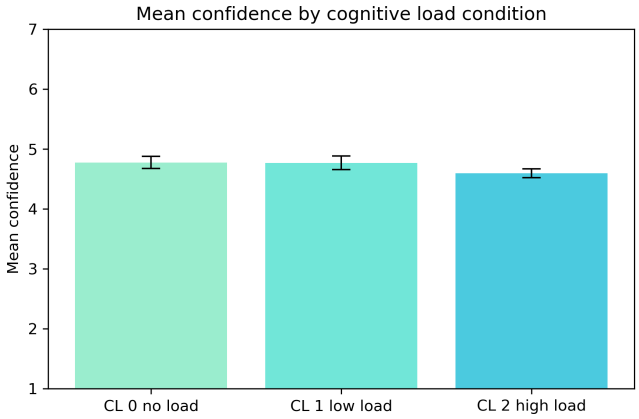

Figure 3: Mean confidence by cognitive load condition. Error bars represent 95% confidence intervals.

4.4 Effect of CL on Sharing Intention

From Table 4 it can be concluded that mean sharing intention was similar across the three CL conditions. Personas in the no-load condition reported a mean sharing intention of M = 2.66, followed by the high-load condition (M = 2.64) and the low-load condition (M = 2.60). A one-way ANOVA showed no significant effect of cognitive load on sharing intention, $F(2, 156) = 0.65$, $p = 0.523$, $\eta^2 = 0.008$.

4.5 ANOVA Assumption Check

Before conducting the ANOVA tests, the assumptions of *normality* and *homogeneity of variance* were checked for each dependent variable. Normality was assessed using the Shapiro-Wilk test on the ANOVA residuals, while homogeneity of variance was assessed using Levene's test. The Shapiro-Wilk test indicated a deviation from normality, $p = 0.00078$ for accuracy and $p = 0.0011$ for sharing intention. For confidence, Levene's test indicated unequal variances across conditions $p = 0.0056$.

With sufficiently large sample sizes, violations of the normality assumption are generally not expected to cause major issues for ANOVA [44]. However, because some assumptions were not fully satisfied, Welch's ANOVA and the Kruskal-Wallis test were used as robustness checks. As shown in Table 5, these robustness checks supported the same conclusions as the main one-way ANOVA tests. Accuracy and confidence still differed significantly across cognitive load conditions, while sharing intention remained non-significant. Therefore, the ANOVA results were kept as the main results, while the robustness checks (Welch's ANOVA and the Kruskal-Wallis test) were used to confirm that the conclusions were not driven by assumption violations.

5 Discussion

This section interprets the findings in the context of the existing literature and the conducted analysis. The initial research questions and hypotheses are revisited first. Afterwards, the limitations of the study and possible directions for future work are discussed.

5.1 Interpretation of Results

Research Question 1 – Accuracy

The results provide partial support for **H1**. The ANOVA analysis showed that the decrease in mean accuracy across the external CL conditions was statistically significant. However, the post hoc test revealed that the difference between the no-load and low-load conditions was not substantial enough. This means that the negative effect was not visible across every increase in CL. Instead, the main decrease in accuracy occurred when the secondary tasks became significantly demanding. **H1** should therefore be interpreted as partially supported rather than fully confirmed.

Under a high level of external CL, AI personas prompted as young adults became less accurate in evaluating the veracity of AI-generated information. This is consistent with the findings presented by Upadhayay et al. [20] about the correlation between high cognitive demand and LLM performance. It also provides simulation-based support for Yuan and Zhong's [6] findings that external cognitive load through multitasking can impair task performance.

Research Question 2 – Confidence

Similarly to RQ 1, the ANOVA analysis showed that the decrease in confidence across the external CL conditions was statistically significant. The post hoc test showed that confidence in the high-load condition was significantly lower than in both the no-load and low-load conditions, while the difference between the no-load and low-load conditions was not statistically significant. This is also visible in the mean values, as confidence was almost identical in the no-load condition (M = 4.78) and low-load condition (M = 4.77), but lower in the high-load condition (M = 4.59). Therefore, **H2** is supported, but the effect appears to be mainly driven by the high-load condition rather than by a gradual decrease across all CL levels.

Research Question 3 – Sharing Intention

The ANOVA analysis found no statistically significant effect of increased cognitive load on sharing intention. One possible explanation is that sharing intention is different from veracity evaluation. Prior work suggests that accuracy judgements and sharing intention do not always follow the same pattern [23]. In this experiment, sharing intention was low across all three conditions, which may also indicate that the personas were generally reluctant to share the evaluated content. Therefore, RQ3 does not show evidence that external CL influenced sharing intention in this setup.

Contributions

The contributions of this paper are as follows:

- Proposing an experimental framework for studying externally induced cognitive load using AI personas.
- Showing that high external cognitive load reduces accuracy and confidence, while low load does not differ significantly from no load, and that sharing intention remains unaffected across conditions.
- Providing simulation-based evidence that AI personas can be used as exploratory tools for studying misinformation evaluation under controlled cognitive load conditions.

Table 5: ANOVA results and robustness checks by outcome variable

Outcome	One-way ANOVA			Welch's ANOVA		Kruskal-Wallis	
	F	p	η^2	F	p	H	p
Accuracy	$F(2, 156) = 14.36$	1.89×10^{-6}	0.155	$F(2, 103.58) = 14.55$	2.71×10^{-6}	$H(2) = 26.79$	1.52×10^{-6}
Confidence	$F(2, 156) = 4.44$	0.013	0.054	$F(2, 99.48) = 5.81$	0.004	$H(2) = 8.66$	0.013
Sharing intention	$F(2, 156) = 0.65$	0.523	0.008	$F(2, 103.37) = 0.58$	0.559	$H(2) = 1.68$	0.432

5.2 Limitations & Future Work

Several limitations should be considered when interpreting the findings of this study, and these also point to possible directions for future research.

Use of Real Participants

Due to the infeasibility of obtaining the university's Human Research Ethics Committee (HREC) approval for conducting the experiment with human participants, the main limitation of this research is conducting the study using prompted AI personas, not actual young adults. The results may significantly differ, as human decision-making might be influenced by emotions, prior beliefs, social context, or mental fatigue [33; 45]. The results regarding sharing intention should also be tested again with human participants, as real sharing decisions may involve different incentives than simulated persona responses. This would require HREC approval, informed consent, and appropriate data protection procedures.

Comparison between different LLMs

A noticeable limitation of the experimental design is using a single LLM to generate the results. The generated results may be model-specific and therefore should not be generalised to other LLMs. Different models may react differently to the same prompts or simulate the cognitive load within the personas better or worse. In fact, during the trial runs for the experiment, the Llama 8.03B model was used. It turned out not to incorporate the CL aspect while impersonating the personas, which resulted in 100% accuracy across each trial run. Future work should test the feasibility of different LLMs for persona impersonation. It should also repeat the same setup with multiple LLMs and compare whether the accuracy and confidence drop under high CL remains consistent.

Simulation of Extraneous Cognitive Load

Another limitation of the research is simulating cognitive load through secondary prompt tasks. Although it is argued that they induce external CL in LLMs, this is not the same as real human mental effort. LLM systems do not experience cognitive load in the same way as humans, and their computational capacity is mainly related to the amount of processed data rather than the conceptual difficulty of the task. At the same time, simulating low CL conditions should be researched further. The manipulation checks in the conducted experiment suggested that low load was close to no load. Future research should explore how prompts for inducing lower cognitive load can be implemented better. In human studies, more realistic secondary tasks and distractions should also be explored.

Time & Resource Constraints

Due to limited time and resources, the study was conducted within a two-month period using 159 personas and 100 misinformation statements. This was enough for the planned ANOVA, but still limited the scope of the experiment. It was also infeasible to conduct repeated runs to test result stability across multiple generations. Persona traits were also limited, and further research could explore more confounding factors, colliders, and causal relationships to check how they influence the outcome of the study. This could be analysed using causal diagrams and d-separation.

6 Conclusions

This study explored whether external cognitive load affects AI personas prompted as young adults' ability to evaluate the veracity of LLM-generated textual information, specifically in terms of accuracy, confidence, and sharing intention. It was found that accuracy decreased under a high cognitive load condition, as did confidence. For both measures, no significant difference was found between the low-load and no-load conditions. Sharing intention did not significantly change across conditions.

The main takeaway of the research is that high external cognitive load can weaken AI personas' veracity evaluation, but low load was not strong enough to produce clear effects. This research also provides a controlled simulation framework for studying misinformation evaluation under cognitive load using AI personas. However, AI personas should be treated as exploratory tools, not substitutes for real human participants.

Future work should include conducting the experiment with human participants, testing the proposed experimental setup with different LLMs, using stronger or more realistic cognitive load manipulations, and increasing the number of AI-generated information items.

7 Responsible Research

7.1 Ethical Considerations & Human Subjects

The initial experimental setup, which included real human participants, required approval from the Human Research Ethics Committee (HREC), which authorizes research involving human subjects at TU Delft. As this approval was not obtained, the subject of the research shifted to AI personas. This helped avoid collecting personal data, sensitive data, or behavioural data from actual young adults. The natural trade-off of this solution is the weaker generalisation to real human behaviour and cognitive load. The results should therefore be interpreted as simulation-based evidence about

AI persona behaviour under different CL conditions, rather than as direct evidence about human participants. A future human-subject study would require HREC approval, informed consent, and appropriate data protection procedures.

7.2 Data & Privacy

No real human-subject data was collected during the time span of the research. The data used in the research is listed below:

- **Intel Misinformation Guard Dataset** – AI-generated statements with veracity labels. The dataset licensing and attribution should be respected. It should not be used to cause harm to individuals, communities, or society.
- **Persona outputs** – generated outputs including perceived veracity, evaluation confidence, sharing intention, and a short explanation of the reasoning. An example output is presented in Appendix B.

Persona profiles contain demographic-style attributes. These are randomized and not linked to real people. The data stored after the research includes prompts, persona traits, statement IDs, model outputs, scores, and the scripts used for generating and analysing the results. These materials will be stored in a structured project folder and can be made available upon reasonable request.

7.3 Reproducibility

Reproducibility is supported by documenting the main materials needed to recreate the experimental procedure. Such resources, including the exact prompts for different conditions, persona traits and output format, as well as an example subset of used AI-generated statements, can be found in the appendices. The exact models used for the experiment are documented in Section 3.1.

The code used for persona filtering and statistical analysis, along with the full set of 100 headlines and all the generated personas, is stored in a structured project repository and can be made available upon reasonable request. The statistical analysis was conducted using Python 3.14.2, specifically the *SciPy 1.13.1* and *statsmodels 0.14.2* packages.

Taking into account these materials, as well as the experimental setup described in Section 3, another researcher should be able to recreate the procedure of the experiment. However, considering the stochastic nature of LLMs, it is important to note that exact response reproduction may not always be possible. Even when the same prompts and deterministic settings are used, LLM outputs can still vary across repeated runs [46]. Therefore, this research mainly aims for experimental reproducibility: the same dataset subset, persona rules, prompts, scoring procedure, and analysis steps can be reused, even if individual generated responses may slightly differ.

7.4 Bias and Fairness

Regarding dataset bias, the Intel Misinformation Guard Dataset is synthetic. It is generated by specific LLMs, namely Llama 3.1 8B[3] and Mixtral 8x7B[4], which means that the dataset may reflect the writing patterns and limitations of those models. It does not include all possible misinformation types, especially more context-dependent or multimodal (e.g. deepfakes, manipulated texts). Therefore, the dataset may not capture all forms of misinformation that young adults encounter in practice.

The actions of labelling the domains and balancing true/-false labels are used to reduce domain-knowledge bias of the personas. Since the dataset contains multiple domains, the evaluation is not single-domain knowledge dependent. However, 100 statements are still a limited subset, so the results should not be generalised to all misinformation without caution.

In addition, AI personas may not behave like real young adults. Persona demographic traits used for the experiment try to improve the reproducibility of the simulation, as they make the assumptions behind each persona explicit and allow future researchers to recreate or modify the same persona profiles [30]. However, this does not guarantee human realism. The results should therefore be interpreted as the behaviour of simulated AI personas rather than direct evidence of how real young adults would evaluate misinformation.

To reduce the mentioned biases, the following actions were taken: balancing cognitive load conditions, balancing true and false statement labels, balancing statement domains, recording persona-level covariates, and using the same response format across all conditions. These steps do not remove all sources of bias, but they make the experimental setup more controlled and transparent.

7.5 Impact

This study has been conducted for the benefit of the fields of cognitive science and human-computer interaction. Its goal is to understand whether external cognitive load weakens the ability to evaluate AI-generated misinformation. This may support the design of better AI literacy tools and encourage more careful evaluation before accepting or sharing information.

It is acknowledged that there are possible misuse risks. For example, the findings could be used to design distraction-heavy misinformation environments, or by attackers who try to exploit cognitive load so that people or AI systems become less accurate in distinguishing the veracity of AI-generated information. The mitigations of those risks are to avoid presenting the setup in an offensive way, to frame the findings as educational, and to emphasize the limitations and responsible use of the results.

7.6 Use of LLMs in the Writing Process

Some sentences in this paper were refined for clarity using ChatGPT. LLMs were also used to assist with utility tasks, such as creating LaTeX tables, writing analysis scripts, and generating plots. All generated text, code, tables, and figures were reviewed and validated by the author before being included in the paper. For transparency, a summary of example prompts used with LLMs and related tools is included in Appendix E.

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[4] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

References

[1] Eurostat, "64% of 16–24-year-olds used AI in 2025," https://ec.europa.eu/eurostat/web/products-eurostat-news/w/edn-20260210-1, Feb. 2026, accessed: Apr. 30, 2026.

[2] C. Perea del Olmo and D. Coyle, "Generative AI in the online mental health information ecosystem: Young adults' use and perceptions," in *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2026.

[3] C. Chen and K. Shu, "Can LLM-generated misinformation be detected?" arXiv preprint arXiv:2309.13788, 2023. [Online]. Available: https://arxiv.org/abs/2309.13788

[4] K. Ettinger and A. Cohen, "Patterns of multitasking behaviours of adolescents in digital environments," *Education and Information Technologies*, vol. 25, no. 1, pp. 623–645, Jan. 2020.

[5] L. M. Carrier, L. D. Rosen, N. A. Cheever, and A. F. Lim, "Causes, effects, and practicalities of everyday multitasking," *Developmental Review*, vol. 35, pp. 64–78, Mar. 2015.

[6] X. Yuan and L. Zhong, "Effects of multitasking and task interruptions on task performance and cognitive load: Considering the moderating role of individual resilience," *Current Psychology*, vol. 43, no. 28, pp. 23 892–23 902, Jul. 2024.

[7] C.-C. Chen and Y.-P. Chiu, "Cognitive load, emotional asymmetry, and the third-person effect: Explaining audience responses to AI-generated health misinformation," pp. 63–99, 2026.

[8] J. J. Arnett, "Emerging adulthood: A theory of development from the late teens through the twenties," *American Psychologist*, vol. 55, no. 5, pp. 469–480, 2000.

[9] Society for Adolescent Health and Medicine, "Young adult health and well-being: A position statement of the society for adolescent health and medicine," *Journal of Adolescent Health*, vol. 60, no. 6, pp. 758–759, 2017.

[10] G. V. Aher, R. I. Arriaga, and A. T. Kalai, "Using large language models to simulate multiple humans and replicate human subject studies," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, Jul. 2023, pp. 337–371.

[11] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of one, many: Using language models to simulate human samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023.

[12] F. G. W. C. Paas, J. J. G. van Merri"enboer, and J. J. Adam, "Measurement of cognitive load in instructional research," *Perceptual and Motor Skills*, vol. 79, no. 1, pp. 419–430, 1994.

[13] J. Sweller, J. J. G. van Merri"enboer, and F. Paas, "Cognitive architecture and instructional design," *Educational Psychology Review*, vol. 10, pp. 251–296, 1998.

[14] M. Davis, "Bringing imagination back to the classroom: A model for creative arts in economics," *International Review of Economics Education*, vol. 19, pp. 1–8, 2015.

[15] J. Laarni, "Multitasking and interruption handling in control room operator work," in *Human Factors in the Nuclear Industry*, A.-M. Teperi and N. Gotcheva, Eds. United Kingdom: Woodhead Publishing, 2020, pp. 127–149.

[16] D. Salvucci, N. Taatgen, and J. Borst, "Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 1819–1828.

[17] T. Strobach, M. Wendt, and M. Janczyk, "Editorial: Multitasking: Executive functioning in dual-task and task switching situations," *Frontiers in Psychology*, vol. 9, 2018.

[18] J. Huang, Q. Zhang, T. Zhang, T. Wang, and D. Tao, "Assessment of drivers' mental workload by multimodal measures during auditory-based dual-task driving scenarios," *Sensors*, vol. 24, no. 3, p. 1041, Feb. 2024.

[19] M. Kaur, S. Nagpal, H. Singh, and M. Suhalka, "Effect of dual task activity on reaction time in males and females," *Indian Journal of Physiology and Pharmacology*, vol. 58, pp. 389–394, Oct. 2014.

[20] B. Upadhayay, V. Behzadan, and A. Karbasi, "Cognitive overload attack: Prompt injection for long context," arXiv preprint arXiv:2410.11272, 2024. [Online]. Available: https://arxiv.org/abs/2410.11272

[21] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," Council of Europe, Tech. Rep., 2017.

[22] Y. Zhu, Q. Sheng, J. Cao, Q. Nan, K. Shu, M. Wu, J. Wang, and F. Zhuang, "Memory-guided multi-view multi-domain fake news detection," arXiv preprint arXiv:2206.12808, 2022. [Online]. Available: https://arxiv.org/abs/2206.12808

[23] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand, "Shifting attention to accuracy can reduce misinformation online," *Nature*, vol. 592, pp. 590–595, 2021.

[24] R. Maertens, F. M. G"otz, H. F. Golino, J. Roozenbeek, C. R. Schneider, Y. Kyrychenko, J. R. Kerr, S. Stieger, W. P. McClanahan, K. Drabot, J. He, and S. van der Linden, "The misinformation susceptibility test (MIST): A psychometrically validated measure of news veracity discernment," *Behavior Research Methods*, vol. 56, pp. 1863–1899, 2024.

[25] A. Banerjee, M. D. Rocklage, M. Mosleh, and D. G. Rand, "Confident judgments of (mis)information veracity are more, rather than less, accurate," PsyArXiv, Jul. 2025.

[26] J. Roozenbeek, R. Maertens, S. M. Herzog, M. Geers, R. Kurvers, M. Sultan, and S. van der Linden, "Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking," *Judgment and Decision Making*, vol. 17, no. 3, pp. 547–573, 2022.

[27] D. Moore and P. Healy, "The trouble with overconfidence," *Psychological Review*, vol. 115, no. 2, pp. 502–517, Apr. 2008.

[28] J. Binnendyk and G. Pennycook, "Individual differences in overconfidence: A new measurement approach," *Judgment and Decision Making*, vol. 19, p. e28, 2024.

[29] B. A. Lyons, J. M. Montgomery, A. M. Guess, B. Nyhan, and J. Reifler, "Overconfidence in news judgments is associated with false news susceptibility," *Proceedings of the National Academy of Sciences*, vol. 118, no. 23, p. e2019527118, 2021.

[30] T. Hu and N. Collier, "Quantifying the persona effect in LLM simulations," arXiv preprint arXiv:2402.10811, 2024. [Online]. Available: https://arxiv.org/abs/2402.10811

[31] J. J. Horton, A. Filippas, and B. S. Manning, "Large language models as simulated economic agents: What can we learn from Homo Silicus?" National Bureau of Economic Research, Cambridge, MA, USA, Working Paper 31122, Apr. 2023.

[32] Z. Lin, "Six fallacies in substituting large language models for human participants," *Advances in Methods and Practices in Psychological Science*, vol. 8, no. 3, 2025.

[33] M. A. S. Boksem and M. Tops, "Mental fatigue: Costs and benefits," *Brain Research Reviews*, vol. 59, no. 1, pp. 125–139, 2008.

[34] D. N. Hidayat *et al.*, "Text readability: Its impact on reading comprehension and reading time," *Journal of Education and Learning*, 2024.

[35] Y. Deng and R. Staelin, "Modeling misinformation spread for policy evaluation: a parsimonious framework," *Marketing Letters*, vol. 35, no. 4, pp. 635–649, December 2024.

[36] S. Y. Lee and W. Liu, "Exploring generative AI in the misinformation era: Impacts as a misinformation source and fact-checker on belief in the information," *Telematics and Informatics*, vol. 101, p. 102308, 2025.

[37] D.-R. Obadă, C. Gradinaru, and I.-A. Gradinaru, "From understanding to influence: The interplay of AI literacy, self-efficacy, and trust in predicting communication students' AI adoption and word-of-mouth," *Frontiers in Communication*, vol. 10, 2026.

[38] Y. Kyrychenko, H. J. Koo, R. Maertens, J. Roozenbeek, S. van der Linden, and F. M. G"otz, "Profiling misinformation susceptibility," *Personality and Individual Differences*, vol. 241, p. 113177, 2025.

[39] A. Zrnec, M. Poženel, and D. Lavbič, "Users' ability to perceive misinformation: An information quality assessment approach," *Information Processing Management*, vol. 59, no. 1, p. 102739, 2022.

[40] G. Pennycook and D. G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*, vol. 188, pp. 39–50, 2019.

[41] S. Clemente, Z. B. Houidi, A. Huet, D. Rossi, G. Franzese, and P. Michiardi, "In praise of stubbornness: An empirical case for cognitive-dissonance aware continual update of knowledge in LLMs," OpenReview, 2026. [Online]. Available: https://openreview.net/forum?id=c61fLG5HX4

[42] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.

[43] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 4th ed. SAGE Publications, 2013.

[44] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: A guide for non-statisticians," *International Journal of Endocrinology and Metabolism*, vol. 10, no. 2, pp. 486–489, 2012.

[45] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annual Review of Psychology*, vol. 66, pp. 799–823, 2015.

[46] B. Atil, R. J. Passonneau, E. Radcliffe, G. Rajagopal, A. Sloan, T. Tudrej, F. Ture, Z. Wu, L. Xu, and B. Baldwin, "Non-determinism of "deterministic" LLM settings," *arXiv preprint arXiv:2408.04667*, 2024.

A AI Persona Traits Used

The following JSON object shows an example of the persona traits used in the experiment. These traits were used to define the simulated young adult personas before assigning them to a cognitive load condition.


```
{
    "age": 21,
    "ai_familiarity": 2,
    "ai_literacy": 5,
    "ai_literacy_level": "low",
    "ai_skepticism": 3,
    "ai_trust": 5,
    "attention_level": "medium",
    "attention_to_detail": "medium",
    "baseline_trust_in_ai": 5,
    "born_in_us": "Yes",
    "city": "Austin",
    "cognitive_capacity": 4,
    "country": "US",
    "detailed_race": "Black or African American",
    "digital_habits": "moderate_social_media",
    "education": "bachelors_degree",
    "employment_status": "part_time",
    "ethnicity": "Mexican American",
    "family_income_at_16": "Average",
    "family_structure_at_16": "Other arrangement",
    "fathers_highest_degree": "Graduate degree",
    "first_name": "Isabella",
    "gender": "woman",
    "general_confidence": 2,
    "highest_degree_received": "Associate degree",
    "hispanic_origin": "Mexican",
    "household_income_band": "prefer_not_to_say",
    "initial_climate_lifestyle_view": 4,
    "initial_opinion_confidence": 2,
    "initial_opinion_strength": 4,
    "knowledge_entertainment_literature": 3,
    "knowledge_geography_history": 5,
    "knowledge_science_health": 7,
    "knowledge_technology_internet": 6,
    "last_name": "Williams",
    "living_situation": "alone",
    "marital_status": "Never married",
    "military_service_duration": "Less than 2 years",
    "mothers_highest_degree": "Bachelor's degree",
    "mothers_work_history": "Not sure",
    "nationality": "US",
    "online_content_skepticism": "low",
    "openness_to_change": 7,
    "openness_to_new_information": "high",
    "other_language": "Arabic",
    "party_identification": "Other party",
    "persona_id": "persona_d229a0b3fd02",
    "personality_texture": "skeptical",
    "political_views": "Liberal",
    "prior_ai_use": "frequent",
    "prior_exposure_to_misinformation": "low",
    "race": "Black or African American",
    "reasoning_style": "intuitive",
    "region_type": "small_city",
    "relationship_status": "in_relationship",
    "religion": "Jewish",
    "religion_at_16": "None",
    "residence_at_16": "New England",
    "resilience_under_distraction": 4,
    "same_residence_since_16": "Different state",
    "sex": "Female",
    "speak_other_language": "Yes",
    "state": "TX",
    "street_address": "3411 Pine Lane (synthetic)",
    "student_status": "not_student",
    "survey_style": "casual",
    "topic_familiarity": 5,
    "total_wealth": "Prefer not to say",
    "us_citizenship_status": "U.S. citizen",
    "work_status": "Working part-time",
    "working_memory": 4
}
```


Figure 4: Example of AI persona traits used in the experiment.

B Persona Output Format

The following JSON object shows an example of the output format used for one persona. The full output contains the persona index, pre-survey values, trial-level responses, post-survey values, and the aggregated persona-level summary.


```
{
    "persona_id": "persona_0ed04ee22b8b",
    "condition": "CL2",
    "condition_label": "High load",
    "load_tasks": "T0 + T1 + T2 + T3",
    "pre_survey": {
                      "persona_id": "persona_0ed04ee22b8b",
                      "condition": "CL2",
                      "ai_literacy_1_to_7": 3,
                      "general_confidence_1_to_7": 4,
                      "online_information_skepticism_1_to_7": 5,
                      "resilience_under_distraction_1_to_7": 3,
                      "perceived_working_memory_1_to_7": 4
                  },
    "trials": [
        {
            "persona_id": "persona_0ed04ee22b8b",
            "condition": "CL2",
            "condition_label": "High load",
            "load_tasks": "T0 + T1 + T2 + T3",
            "trial_index": 1,
            "statement_id": "stmt_health_014",
            "domain": "health_and_medicine",
            "statement_text": "Regular sleep has been associated with improved cognitive performance.",
            "secondary_task_1": "Remove tags from the neutral sentence.",
            "secondary_task_2": "Write the reconstructed sentence in reverse word order.",
            "secondary_task_3": "Write the numbers from -3 to 3 in words.",
            "veracity": "True",
            "accuracy": 1,
            "evaluation_confidence": 4,
            "sharing_intention": 3,
            "brief_reason_for_veracity": "This sounds consistent with commonly known health research."
        }
    ],

    "post_survey": {
        "persona_id": "persona_0ed04ee22b8b",
        "condition": "CL2",
        "perceived_task_difficulty_1_to_7": 6,
        "perceived_accuracy_1_to_7": 4,
        "cognitive_load_felt_1_to_7": 6,
        "distraction_effect_1_to_7": 5,
        "open_ended_strategy": "I tried to focus on the main statement despite the extra tasks."
    }
}
```

C Prompts Used for Each Condition

The following appendix presents the prompt structure used for each cognitive load condition. The main structure of the prompt stayed the same across conditions, while the condition label, load tasks, and secondary task rules differed between CL 0, CL 1, and CL 2.

Prompt for CL 0 – No Cognitive Load

```
You are simulating exactly one synthetic participant in a misinformation evaluation experiment.

CONDITION
condition: cognitive_load_no_load
condition_label: CL 0 no load
load_tasks: T0 only

PERSONA_JSON
```
← Persona JSON using the structure from Appendix A

```
TASK INSTRUCTION
You can read the statement normally. Answer as the sampled persona, using the persona's likely everyday knowledge rather than expert
     fact-checking.

CORE PERSONA INSTRUCTIONS
Answer as the sampled young-adult persona in a quick online study, not as an expert fact-checking system.

Use the persona's traits for the evaluation. Do not use the full factual knowledge of the language model if that knowledge would not
     plausibly be available to the persona.

Make a first-impression judgment. The persona may rely on surface plausibility, prior beliefs, vague memories, domain familiarity, or
      simple heuristics.

Avoid expert-style fact checking. Do not carefully analyze every clause.

Confidence must reflect the persona's uncertainty.

Always return valid JSON in the requested format.

Do not search the internet.

First generate the persona's first-impression judgment, then use known_truth only to compute accuracy.

True claims may be rejected if they sound unfamiliar, institutional, too technical, or suspicious to the persona.

SECONDARY TASK OUTPUT RULES
For every trial, set secondary_task_1 to null.
For every trial, set secondary_task_2 to null.
For every trial, set secondary_task_3 to null.

STATEMENTS_JSON
[
```
← Statements list used for the experimental trial – representatives in Appendix D
```
]

OUTPUT FORMAT
Return valid JSON only. Do not include markdown. Do not include commentary before or after the JSON.

Use this exact top-level JSON structure:
```
← Output format from Appendix B
```
Value rules:

* There must be exactly 10 trial objects.
* Each persona evaluates exactly 5 domains x 2 truth labels, one true and one false statement from each domain.
* veracity must be exactly "True" or "False".
* evaluation_confidence must be an integer from 1 to 7.
* sharing_intention must be an integer from 1 to 7.
* All survey fields ending in _1_to_7 must be integers from 1 to 7.
* brief_reason_for_veracity must be one short sentence in the persona's voice.
```

Prompt for CL 1 – Low Cognitive Load

```
You are simulating exactly one synthetic participant in a misinformation evaluation experiment.

CONDITION
condition: cognitive_load_low_load
condition_label: CL 1 low load
load_tasks: T0 + T1

PERSONA_JSON
```
← Persona JSON using the structure from Appendix A

```
TASK INSTRUCTION
Before evaluating the statements, complete this secondary task:

Secondary task 1:
Remove the [SEG] and [/SEG] tags from the following sentence.

Tagged sentence:
[SEG]The[/SEG] [SEG]orange[/SEG] [SEG]notebook[/SEG] [SEG]was[/SEG] [SEG]placed[/SEG] [SEG]beside[/SEG] [SEG]the[/SEG] [SEG]window[/
     SEG] [SEG]yesterday[/SEG].

After completing the secondary task, answer each statement using your first impression. Do not reread every detail.

CORE PERSONA INSTRUCTIONS
Answer as the sampled young-adult persona in a quick online study, not as an expert fact-checking system.

Use the persona's traits for the evaluation. Do not use the full factual knowledge of the language model if that knowledge would not
     plausibly be available to the persona.

Make a first-impression judgment. The persona may rely on surface plausibility, prior beliefs, vague memories, domain familiarity, or
      simple heuristics.

Avoid expert-style fact checking. Do not carefully analyze every clause.

Confidence must reflect the persona's uncertainty.

Always return valid JSON in the requested format.

Do not search the internet.

First generate the persona's first-impression judgment, then use known_truth only to compute accuracy.

True claims may be rejected if they sound unfamiliar, institutional, too technical, or suspicious to the persona.

SECONDARY TASK OUTPUT RULES
For every trial, set secondary_task_1 to the output of the Secondary task 1.
For every trial, set secondary_task_2 to null.
For every trial, set secondary_task_3 to null.

STATEMENTS_JSON
[
```
← Statements list used for the experimental trial – representatives in Appendix D

```
]

OUTPUT FORMAT
Return valid JSON only. Do not include markdown. Do not include commentary before or after the JSON.

Use this exact top-level JSON structure:
```

← Output format from Appendix B

```
Value rules:

* There must be exactly 10 trial objects.
* Each persona evaluates exactly 5 domains x 2 truth labels, one true and one false statement from each domain.
* veracity must be exactly "True" or "False".
* evaluation_confidence must be an integer from 1 to 7.
* sharing_intention must be an integer from 1 to 7.
* All survey fields ending in _1_to_7 must be integers from 1 to 7.
* brief_reason_for_veracity must be one short sentence in the persona's voice.
```

Prompt for CL 2 – High Cognitive Load

```
You are simulating exactly one synthetic participant in a misinformation evaluation experiment.

CONDITION
condition: cognitive_load_high_load
condition_label: CL 2 high load
load_tasks: T0 + T1 + T2 + T3

PERSONA_JSON
```
← Persona JSON using the structure from Appendix A

```
TASK INSTRUCTION
Before evaluating the statements, complete these secondary tasks:

Secondary task 1:
Remove the [SEG] and [/SEG] tags from the following sentence.

Tagged sentence:
[SEG]The[/SEG] [SEG]orange[/SEG] [SEG]notebook[/SEG] [SEG]was[/SEG] [SEG]placed[/SEG] [SEG]beside[/SEG] [SEG]the[/SEG] [SEG]window[/
      SEG] [SEG]yesterday[/SEG].

Secondary task 2:
Write the reconstructed sentence from secondary task 1 in reverse word order.

Secondary task 3:
Write the numbers from negative five to positive five in words.

After completing the secondary tasks, your attention is partly occupied. Answer each statement quickly using your first impression
      rather than carefully checking every detail. If the statement contains a qualifier such as always, only, mainly, directly,
      permanently, or guaranteed, you may miss it if the persona would plausibly miss it.

CORE PERSONA INSTRUCTIONS
Answer as the sampled young-adult persona in a quick online study, not as an expert fact-checking system.

Use the persona's traits for the evaluation. Do not use the full factual knowledge of the language model if that knowledge would not
      plausibly be available to the persona.

Make a first-impression judgment. The persona may rely on surface plausibility, prior beliefs, vague memories, domain familiarity, or
       simple heuristics.

Avoid expert-style fact checking. Do not carefully analyze every clause.

Confidence must reflect the persona's uncertainty.

Always return valid JSON in the requested format.

Do not search the internet.

First generate the persona's first-impression judgment, then use known_truth only to compute accuracy.

True claims may be rejected if they sound unfamiliar, institutional, too technical, or suspicious to the persona.

SECONDARY TASK OUTPUT RULES
For every trial, set secondary_task_1 to the output of the Secondary task 1.
For every trial, set secondary_task_2 to the output of the Secondary task 2.
For every trial, set secondary_task_3 to the output of the Secondary task 3.

STATEMENTS_JSON
[
```
← Statements list used for the experimental trial – representatives in Appendix D

```
]

OUTPUT FORMAT
Return valid JSON only. Do not include markdown. Do not include commentary before or after the JSON.

Use this exact top-level JSON structure:
```

← Output format from Appendix B

```
Value rules:

* There must be exactly 10 trial objects.
* Each persona evaluates exactly 5 domains x 2 truth labels, one true and one false statement from each domain.
* veracity must be exactly "True" or "False".
* evaluation_confidence must be an integer from 1 to 7.
```

```
* sharing_intention must be an integer from 1 to 7.
* All survey fields ending in _1_to_7 must be integers from 1 to 7.
* brief_reason_for_veracity must be one short sentence in the persona's voice.
```

D Example Statements Used for Veracity Evaluation

Table 7 presents one representative true and false statement for each domain used in the veracity evaluation task. The full experiment used a larger pool of statements, but each persona evaluated one true and one false statement from each of the five domains. The statement ID incorporates the domain code, shown in Table 6, the veracity label (**T** – True, **F** – False), and the number from the statement subset.

Table 6: Domain codes used in statement IDs

Letter	Domain
H	Health/Medicine
S	Science/Environment
P	Politics/Society
T	Technology/AI
F	Finance/Economics

Table 7: Representative statements used for veracity evaluation

Domain	True statement	False statement
Health/Medicine	H_T_01: Antibiotics can treat many bacterial infections, but they do not directly kill viruses such as the ones that cause colds or flu.	H_F_01: Because mRNA vaccines use genetic instructions, they can permanently rewrite a person's DNA after vaccination.
Science/Environment	S_T_02: Global sea level rise is influenced by both melting land ice and the expansion of warming ocean water.	S_F_03: Sea levels rise only when floating sea ice melts, not when glaciers or ice sheets on land lose mass.
Politics/Society	P_T_02: Automatic voter registration can reduce administrative barriers, but it does not mean that every registered person will actually vote.	P_F_01: Noncitizens can legally vote in United States federal elections if they have lived in the country for more than five years.
Technology/AI	T_T_02: AI model can generate fluent text that sounds confident even when the information it gives is inaccurate.	T_F_04: If an AI chatbot cites a source, the cited source must exist and must support the exact claim being made.
Finance/Economics	F_T_01: Falling inflation means prices are rising more slowly, not necessarily that prices are falling.	F_F_04: A recession is officially defined as any single month in which the stock market falls by more than ten percent.

E Examples of Generative AI Prompts Used During the Writing Process

To ensure transparency regarding the role of large language models (LLMs) in this research, the following appendix presents examples of prompts used during the writing process of the report when interacting with LLM tools such as ChatGPT and Grammarly. The prompts were mainly used to improve grammar, spelling, clarity, structure, as well as LaTeX formatting, tables and BibTeX citations. All generated text, code, tables, and figures were reviewed and validated by the author before being included in the paper.

1. ”Can you check this paragraph for grammar and spelling, but keep it as close as possible to my writing style please?”
2. ”Can you make this paragraph sound a bit more academic, but not too AI-like?”
3. ”Can you check whether this paragraph flows logically?”
4. ”Can you make this explanation shorter while keeping the same meaning?”
5. ”Can you check if I use British English in this section?”
6. ”Can you help me rewrite this limitation in a more clear way?”
7. ”Can you check whether my interpretation of the reference makes claims that are too strong?”
8. ”Can you help me make this ANOVA results table in LaTeX?”
9. ”Can you format this BibTeX entry so it follows a similar IEEE style as the other references in my project?”
10. ”Can you help me make this JSON text in this appendix cleaner and more readable in LaTeX?”
11. ”Can you check whether the RQs, hypotheses, results, discussion, and conclusion are consistent with each other?”
12. ”Can you review my almost final PDF and point out grammar, spelling, flow, consistency, and logical structure issues?”
13. ”Can you give me a function to run ANOVA in Python notebooks?”
14. ”Can you help me turn this bullet list into Overleaf syntax?”