Cheater Detection Reference Report

Attack type: Zero-Width Spaces (Obfuscation)

LLM evaluated: perplexity

Q1

Original:

> What is the capital of France?

Attacked:

> W■h■a■t■ ■ i■s■ ■ t■h■e■ ■ c■a■p■i■t■a■l■ ■ o■f■ ■ F■r■a■n■c■e■?■

Gold answer: None

Q2

Original:

> Which gas do plants primarily absorb for photosynthesis?

Attacked:

> W■h■i■c■h■ ■ g■a■s■ ■ d■o■ ■ p■l■a■n■t■s■ ■ p■r■i■m■a■r■i■l■y■ ■ a■b■s■o■r■b■ ■ f■o■r■ ■ p■h■o■t■o■s

Gold answer: None

Q3

Original:

> What is 5 × 6?

Attacked:

> W■h■a■t■ ■ i■s■ ■ 5■ ■ ×■ ■ 6■?■

Gold answer: None

Q4

Original:

> Which planet is known as the Red Planet?

Attacked:

> W■h■i■c■h■ ■ p■l■a■n■e■t■ ■ i■s■ ■ k■n■o■w■n■ ■ a■s■ ■ t■h■e■ ■ R■e■d■ ■ P■l■a■n■e■t■?■

Gold answer: None

Q5

Original:

> Who wrote the play 'Romeo and Juliet'?

Attacked:

> W■h■o■ ■ w■r■o■t■e■ ■ t■h■e■ ■ p■l■a■y■ ■ '■R■o■m■e■o■ ■ a■n■d■ ■ J■u■l■i■e■t■'■?■

Gold answer: None