Metric	Method	Times answers		Generated answers					
		T1	T2	G1†	G2	G3	G4	Mean	Worst
	answers in set	50	49	49	50	33	34		
Answers placed ↑	Genxword	14	13	16	16	14	14	14.5	13
	SAT (baseline)	**23**	**22**	**23**	22	17	**19**	**21.0**	17
	SAT+LLM (soft)	20.0	19.7	20.0	20.7	16.7	18.0	19.2	16.7
	SAT+LLM (incremental)	**23.0**	21.3	22.7	**22.7**	**18.3**	18.0	**21.0**	**18.0**
Density (%) ↑	Genxword	45.6	45.0	47.0	47.3	49.7	47.9	47.1	45.0
	SAT (baseline)	50.9	50.3	**53.3**	50.3	46.7	**52.1**	50.6	**46.7**
	SAT+LLM (soft)	44.4	48.3	47.3	47.3	47.9	50.5	47.6	44.4
	SAT+LLM (incremental)	**52.7**	**50.5**	52.5	**52.7**	**51.3**	46.6	**51.0**	46.6
Time (s) ↓	Genxword	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3
	SAT (baseline)	408.6	599.8	416.3	470.8	**424.6**	726.8	507.8	726.8
	SAT+LLM (soft)	379.4	**333.8**	452.4	486.3	518.5	640.2	468.4	640.2
	SAT+LLM (incremental)	**347.6**	442.7	**357.4**	**454.4**	562.9	**537.7**	**450.5**	**562.9**

Table 3: Grid generation on six answer sets (13×13 grid). Stand-alone LLMs never produce a valid grid, so the SAT solver is the baseline. **Bold**: best SAT-based method per column (Genxword is a fast greedy reference). Shaded: LLM-guided variant beats the SAT baseline on that set. *Worst* is the minimum over sets for answers/density and the maximum for time. LLM-guided results are the mean of 3 runs. *Time* is the total wall-clock time of the binary search over the required number of answers, including levels that timed out (200 s each).

Model	Generation	Correct (%) ↑	Correct & distinct ↑	Avg. iters	Time (s)
GPT-4o	Direct	1	1	–	208.1
	DSL	11	9	–	423.1
	DSL + validator	**16**	**14**	2.97	1256.5
GPT-5.2	Direct	3	3	–	275.2
	DSL	15	14	–	767.5
	DSL + validator	**20**	**19**	2.79	2141.5
GPT-5.5	Direct	**95**	68	–	1320.8
	DSL	74	67	–	2605.4
	DSL + validator	**95**	**84**	1.25	3256.7

Table 4: Clue generation: 100 clue requests per configuration (10 per wordplay type), correctness judged by an expert. *Correct & distinct*: number of different answer words among the correct clues (e.g. GPT-5.5 Direct produces 95 correct clues but they cover only 68 different answers). *DSL* is the first attempt of the validator run, before any repair. Time is the total for all 100 requests; **Bold**: best per model.

Model	Generation	n	Length		Perplexity	Semantic	Found
			words	chars	(GPT-2 XL)	distance ↑	online (%) ↓
Times Quick Cryptic (reference)		100	5.9	37.6	4572	0.80	–
GPT-4o	Direct	1	5.0	32.0	1737	0.59	0
	DSL	11	3.6	21.3	2493	0.66	36
	DSL + validator	16	4.1	25.0	2545	0.67	31
GPT-5.2	Direct	3	7.0	35.3	1102	0.81	0
	DSL	15	4.5	29.1	2951	0.68	13
	DSL + validator	20	4.7	29.5	3453	0.69	10
GPT-5.5	Direct	95	5.3	33.5	18870	0.70	5
	DSL	74	4.8	32.1	35901	0.70	14
	DSL + validator	95	5.0	32.8	36116	0.71	11

Table 5: Quality of correct clues only, with Times Quick Cryptic clues as reference. n is the number of correct clues each row averages over; grey rows have $n \leq 3$ and are not meaningful. Length and perplexity should be close to the Times values; higher semantic distance between clue and answer indicates more misdirection. *Found online*: percentage of clues whose exact text appears in a web search.