Model	Total Score	Instances Covered	Average Score
gemini-2.0-flash	53.748	67	0.802
gemini-2.5-flash	69.426	86	0.807
gemini-2.5-flash-lite	69.040	85	0.812
gemini-2.5-pro	41.594	51	0.815
allam-2-7b	0.0	10	0.0
groq/compound	8.246	9	0.916
groq/compound-mini	29.623	35	0.846
llama-3.1-8b-instant	56.228	72	0.780
llama-3.3-70b-versatile	9.496	10	0.949
meta-llama/llama-4-maverick-17b-128e-instruct	63.548	72	0.882
meta-llama/llama-4-scout-17b-16e-instruct	57.814	69	0.837
meta-llama/llama-guard-4-12b	0.0	77	0.0
moonshotai/kimi-k2-instruct	65.816	75	0.877
moonshotai/kimi-k2-instruct-0905	66.186	75	0.882
openai/gpt-oss-120b	64.508	74	0.871
openai/gpt-oss-20b	63.328	75	0.844
qwen/qwen3-32b	48.018	65	0.738

Table 1: Initial tests giving plain scripts to all the LLMs, In this table: column "Model" contains the names of each tested LLM, " Total Score" ris the equivalent of "Parallel Score" as explained in **??**, "Instances Covered" contains the amount of instances that gave a viable result so the ones that did not exceed the token limitations portrayed in **??** and **??**, "Average Score" represents $\frac{TotalScore}{InstancesCovered}$ to show the solver performance in the evaluated instances.

Model	Total Score	Instances Covered	Average Score
gemini-2.5-flash-lite	64.363	85	0.794
gemini-2.5-flash	60.962	85	0.734
moonshotai/kimi-k2-instruct-0905	59.680	75	0.817
moonshotai/kimi-k2-instruct	58.609	75	0.837
openai/gpt-oss-120b	58.166	74	0.796
openai/gpt-oss-20b	57.154	75	0.828
meta-llama/llama-4-maverick-17b-128e-instruct	56.297	72	0.804
meta-llama/llama-4-scout-17b-16e-instruct	54.305	69	0.798
gemini-2.0-flash	43.412	67	0.700
qwen/qwen3-32b	42.116	65	0.779
gemini-2.5-pro	37.640	51	0.738
llama-3.1-8b-instant	36.082	72	0.546
groq/compound-mini	25.422	35	0.726
llama-3.3-70b-versatile	7.454	10	0.745
groq/compound	5.197	9	0.577
allam-2-7b	0.0	4	0.0

Table 2: Initial tests giving plain scripts to all the LLMs, in this table "Total Score" is the equivalent of "Single Score" as explained in **??**. "Instances Covered" and "Average Score" are calculated the same way as in **??**

(a) Scores obtained by all of the considered LLM from parallel-solver evaluation, "Total Score" is the equivalent of "Single Score" as explained in **??**.

(b) Scores obtained by all of the considered LLM from single-solver evaluation, " Total Score" ris the equivalent of "Parallel Score" as explained in **??**

Figure 1: Histograms displaying performances of All the LLMs with plain scripts

model	InstancesCovered	AS	SBS	VBS	ClosedGap
gemini-2.5-flash-lite	85	64.363	76.964	89.000	-1.047
gemini-2.5-flash	85	60.962	76.964	89.000	-1.330
moonshotai/kimi-k2-instruct-0905	75	59.680	76.964	89.000	-1.436
moonshotai/kimi-k2-instruct	75	58.609	76.964	89.000	-1.525
openai/gpt-oss-120b	74	58.166	76.964	89.000	-1.562
openai/gpt-oss-20b	75	57.154	76.964	89.000	-1.646
meta-llama/llama-4-maverick-17b-128e-instruct	72	56.297	76.964	89.000	-1.717
meta-llama/llama-4-scout-17b-16e-instruct	69	54.305	76.964	89.000	-1.883
gemini-2.0-flash	67	43.413	76.964	89.000	-2.788
qwen/qwen3-32b	65	42.117	76.964	89.000	-2.895
gemini-2.5-pro	51	37.641	76.964	89.000	-3.267
llama-3.1-8b-instant	72	36.082	76.964	89.000	-3.397
groq/compound-mini	35	25.423	76.964	89.000	-4.282
llama-3.3-70b-versatile	10	7.455	76.964	89.000	-5.775
groq/compound	9	5.197	76.964	89.000	-5.963

Table 3: Initial tests giving plain scripts to all the LLMs, in this table: "Instances Covered" is the same as in **??**, "AS" is the same as the "Single Score" explained in **??**, "SBS" displays the sum of all the scores obtained by the single best solver (namely, `or-tools_cp-sat-free`) in every instance, and "VBS" displays the score obtained with the use of an ipothetical virtual best solver, giving the maximum obtainable score on every instance. Finally, "Closed Gap" is calculated as in **??**

Model	Total Score	Instances Covered	Average Score
gemini-2.5-flash	79.105	100	0.791
gemini-2.5-flash-lite	80.740	100	0.807
moonshotai/kimi-k2-instruct	83.268	100	0.832
moonshotai/kimi-k2-instruct-0905	82.656	100	0.826
openai/gpt-oss-120b	82.226	100	0.822

Table 4: Tests on sanitized scripts given to the 5 best performing LLMs, parallel-solver evaluation. Columns are calculated as in **??**

Model	Total Score	Instances Covered	Average Score
openai/gpt-oss-120b	74.488	100	0.744
moonshotai/kimi-k2-instruct-0905	71.622	100	0.753
moonshotai/kimi-k2-instruct	70.939	100	0.723
gemini-2.5-flash-lite	70.145	100	0.738
gemini-2.5-flash	69.763	98	0.742

Table 5: Tests on sanitized scripts given to the 5 best performing LLMs, single-solver evaluation. Columns are calculated as in **??**

Model	Instances Covered	AS	SBS	VBS	Closed Gap
openai/gpt-oss-120b	100	74.488	76.964	89.0	-0.205
moonshotai/kimi-k2-instruct-0905	100	71.622	76.964	89.0	-0.443
moonshotai/kimi-k2-instruct	100	70.939	76.964	89.0	-0.500
gemini-2.5-flash-lite	100	70.145	76.964	89.0	-0.566
gemini-2.5-flash	98	69.763	76.964	89.0	-0.598

Table 6: Tests on sanitized scripts given to the 5 best performing LLMs, closed gap. Columns are calculated as in **??**

Model	Total Score	Instances Covered	Average Score
gemini-2.5-flash	59.154	75	0.788
gemini-2.5-flash-lite	82.620	100	0.826
moonshotai/kimi-k2-instruct	81.889	100	0.818
moonshotai/kimi-k2-instruct-0905	83.182	100	0.831
openai/gpt-oss-120b	83.249	100	0.832

Table 7: Test on sanitized scripts combined with textual problem description, parallel-solver evaluation. Columns are calculated as in **??**

Model	Total Score	Instances Covered	Average Score
moonshotai/kimi-k2-instruct-0905	72.605	100	0.748
openai/gpt-oss-120b	70.740	100	0.721
gemini-2.5-flash-lite	69.042	100	0.719
moonshotai/kimi-k2-instruct	67.554	100	0.718
gemini-2.5-flash	49.896	73	0.723

Table 8: Test on sanitized scripts combined with textual problem description, single-solver evaluation. Columns are calculated as in **??**

Model	Instances Covered	AS	SBS	VBS	Closed Gap
moonshotai/kimi-k2-instruct-0905	100	72.605	76.964	89.0	-0.362
openai/gpt-oss-120b	100	70.740	76.964	89.0	-0.517
gemini-2.5-flash-lite	100	69.042	76.964	89.0	-0.658
moonshotai/kimi-k2-instruct	100	67.554	76.964	89.0	-0.781
gemini-2.5-flash	73	49.896	76.964	89.0	-2.248

Table 9: Test on sanitized scripts combined with textual problem description, closed gap. Columns are calculated as in **??**

Model	Total Score	Instances Covered	Average Score
gemini-2.5-flash	83.137	100	0.831
gemini-2.5-flash-lite	77.746	100	0.777
moonshotai/kimi-k2-instruct	82.236	100	0.822
moonshotai/kimi-k2-instruct-0905	82.488	100	0.824
openai/gpt-oss-120b	78.799	100	0.787

Table 10: Test with sanitized scripts combined with solvers description in a multi turn setup, parallel-solver evaluation. Columns are calculated as in **??**

Model	Total Score	Instances Covered	Average Score	
moonshotai/kimi-k2-instruct	76.964	100	0.769	
moonshotai/kimi-k2-instruct-0905	76.964	100	0.769	
gemni	gemini-2.5-flash	71.686	100	0.716
openai/gpt-oss-120b	70.974	100	0.716	
gemni	gemini-2.5-flash-lite	54.713	100	0.552

Table 11: Test with sanitized scripts combined with solvers description in a multi turn setup, single-solver evaluation. Columns are calculated as in **??**

Model	Instances Covered	AS	SBS	VBS	Closed Gap
moonshotai/kimi-k2-instruct	100	76.964	76.964	89.0	0.0
moonshotai/kimi-k2-instruct-0905	100	76.964	76.964	89.0	0.0
gemini-2.5-flash	100	71.686	76.964	89.0	-0.438
openai/gpt-oss-120b	100	70.974	76.964	89.0	-0.497
gemini-2.5-flash-lite	100	54.713	76.964	89.0	-1.848

Table 12: Test with sanitized scripts combined with solvers description in a multi turn setup, closed gap. Columns are calculated as in **??**

Model	Total Score	Instances Covered	Average Score
gemini-2.5-flash	83.650	100	0.836
gemini-2.5-flash-lite	78.147	100	0.781
moonshotai/kimi-k2-instruct	80.417	99	0.812
moonshotai/kimi-k2-instruct-0905	82.218	100	0.822
openai/gpt-oss-120b	80.295	100	0.802

Table 13: Test with sanitized scripts combined with both solvers description, and problem description in a multi turn setup, parallel-solver evaluation. Columns are calculated as in **??**

Model	Total Score	Instances Covered	Average Score
openai/gpt-oss-120b	77.260	100	0.780
moonshotai/kimi-k2-instruct-0905	76.964	100	0.769
moonshotai/kimi-k2-instruct	74.464	99	0.752
gemini-2.5-flash	73.551	100	0.750
gemini-2.5-flash-lite	63.062	100	0.630

Table 14: Test with sanitized scripts combined with both solvers description, and problem description in a multi turn setup, single-solver evaluation. Columns are calculated as in **??**

Model	Instances Covered	AS	SBS	VBS	Closed Gap
openai/gpt-oss-120b	100	77.260	76.964	89.0	0.024
moonshotai/kimi-k2-instruct-0905	100	76.964	76.964	89.0	0.0
moonshotai/kimi-k2-instruct	99	74.464	76.964	89.0	-0.207
gemini-2.5-flash	100	73.551	76.964	89.0	-0.283
gemini-2.5-flash-lite	100	63.062	76.964	89.0	-1.155

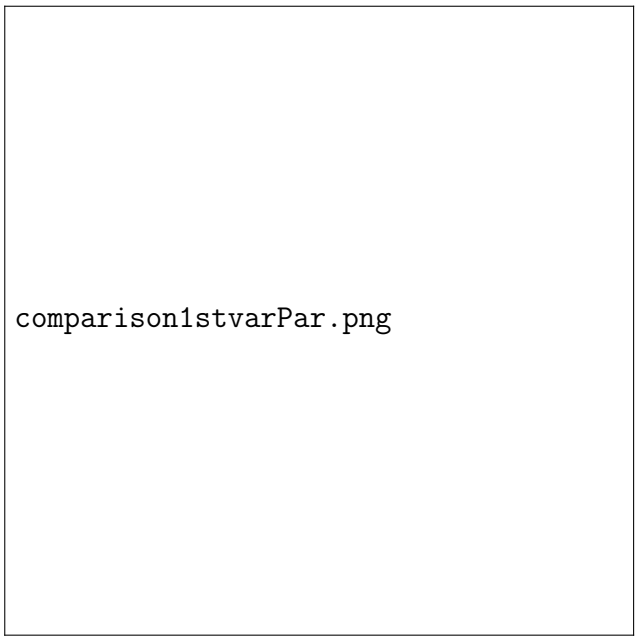
Table 15: Test with sanitized scripts combined with both solvers description, and problem description in a multi turn setup, closed gap. Columns are calculated as in **??**

Type	Solver	Total Score
Solver	or-tools_cp-sat-par	88.117
LLM Variant	gemini-2.5-flash (Solvers Description + Problem Description)	83.650
LLM Variant	moonshotai/kimi-k2-instruct	83.268
LLM Variant	openai/gpt-oss-120b (Problem Description)	83.249
LLM Variant	moonshotai/kimi-k2-instruct-0905 (Problem Description)	83.182
LLM Variant	gemini-2.5-flash (Solvers Description)	83.137
LLM Variant	moonshotai/kimi-k2-instruct-0905	82.656
LLM Variant	gemini-2.5-flash-lite (Problem Description)	82.620
LLM Variant	moonshotai/kimi-k2-instruct-0905 (Solvers Description)	82.488
LLM Variant	moonshotai/kimi-k2-instruct (Solvers Description)	82.236
LLM Variant	openai/gpt-oss-120b	82.226
LLM Variant	moonshotai/kimi-k2-instruct-0905 (Solvers Description + Problem Description)	82.218
LLM Variant	moonshotai/kimi-k2-instruct (Problem Description)	81.889
LLM Variant	gemini-2.5-flash-lite	80.740
LLM Variant	openai/gpt-oss-120b (Solvers Description + Problem Description)	80.295
LLM Variant	gemini-2.5-flash	79.105
LLM Variant	openai/gpt-oss-120b (Solvers Description)	78.799
LLM Variant	gemini-2.5-flash-lite (Solvers Description + Problem Description)	78.147
LLM Variant	gemini-2.5-flash-lite (Solvers Description)	77.746
Solver	chuffed-free	74.819
Solver	picatsat-free	70.647
Solver	huub-free	68.784
Solver	gurobi-par	61.081
Solver	cplex-par	60.301
Solver	choco-solver__cp_-par	59.365
Solver	izplus-par	58.216
Solver	pumpkin-free	57.681
Solver	choco-solver__cp-sat_-par	56.896
Solver	gecode-par	54.283
Solver	cp_optimizer-par	53.877
Solver	gecode_dexter-open	47.304
Solver	or-tools_cp-sat_ls-par	46.292
Solver	jacop-free	44.373
Solver	sicstus_prolog-free	43.548
Solver	yuck-par	38.321
Solver	scip-par	36.675
Solver	highs-par	33.598
Solver	cbc-par	26.334
Solver	atlantis-free	1.555

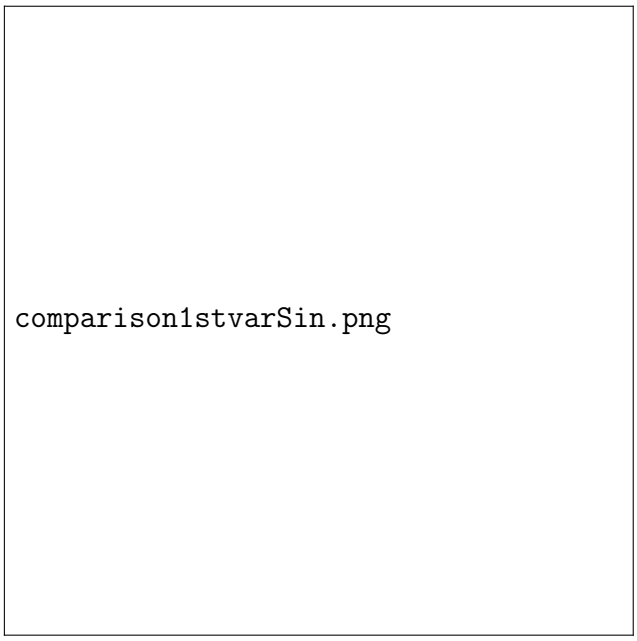
Table 16: Table displaying all the different LLM variants results from parallel-solver evaluation, "Total Score" is calculated as "Parallel Score" in **??**, and compared to all of the single solvers in open category of the MiniZinc Challenge [**?**]

Type	Solver	TotalScore
LLM Variant	**openai/gpt-oss-120b (Solvers Description + Problem Description)**	**77.260**
LLM Variant	moonshotai/kimi-k2-instruct-0905 (Solvers Description + Problem Description)	76.964
LLM Variant	moonshotai/kimi-k2-instruct-0905 (Solvers Description)	76.964
LLM Variant	moonshotai/kimi-k2-instruct (Solvers Description)	76.964
Solver	or-tools_cp-sat-free	76.964
LLM Variant	openai/gpt-oss-120b (Simple)	74.488
LLM Variant	moonshotai/kimi-k2-instruct (Solvers Description + Problem Description)	74.464
Solver	chuffed-free	74.456
LLM Variant	gemini-2.5-flash (Solvers Description + Problem Description)	73.551
LLM Variant	moonshotai/kimi-k2-instruct-0905 (Problem Description)	72.605
LLM Variant	gemini-2.5-flash (Solvers Description)	71.686
LLM Variant	moonshotai/kimi-k2-instruct-0905 (Simple)	71.622
LLM Variant	openai/gpt-oss-120b (Solvers Description)	70.974
LLM Variant	moonshotai/kimi-k2-instruct (Simple)	70.939
Solver	picatsat-free	70.933
LLM Variant	openai/gpt-oss-120b (Problem Description)	70.740
LLM Variant	gemini-2.5-flash-lite (Simple)	70.145
LLM Variant	gemini-2.5-flash (Simple)	69.763
LLM Variant	gemini-2.5-flash-lite (Problem Description)	69.042
Solver	huub-free	68.497
LLM Variant	moonshotai/kimi-k2-instruct (Problem Description)	67.554
LLM Variant	gemini-2.5-flash-lite (Solvers Description + Problem Description)	63.062
Solver	choco-solver__cp-sat_-free	60.808
Solver	izplus-free	58.672
Solver	pumpkin-free	58.543
Solver	choco-solver__cp_-free	58.404
Solver	gurobi-free	55.384
LLM Variant	gemini-2.5-flash-lite (Solvers Description)	54.713
Solver	cp_optimizer-free	50.992
Solver	gecode-fd	50.073
LLM Variant	gemini-2.5-flash (Problem Description)	49.896
Solver	cplex-free	48.514
Solver	sicstus_prolog-free	44.592
Solver	jacop-free	44.549
Solver	or-tools_cp-sat_ls-free	43.222
Solver	scip-free	36.902
Solver	yuck-free	34.715
Solver	highs-free	34.418
Solver	cbc-free	22.615
Solver	atlantis-free	1.5

Table 17: Table displaying all the first LLM variants results given single-solver evaluation, "Total Score" is calculated as "Single Score" in **??**, and compared to all of the single solvers in free category of the MiniZinc Challenge [**?**]



(a) Histograms to visualize the difference between first LLM variants and single solvers from open category performance, "Total Score" is the equivalent of "Parallel Score" as explained in **??**.



(b) Histograms to visualize the difference between first LLM variants and single solvers free category performance, "Total Score" is the equivalent of "Single Score" as explained in **??**.

Figure 2: Histograms to visualize the difference between first LLM variants and single solvers

Category	Number	Description (summary)
Variables	27	Counts of variables (including constants, aliases, defined and introduced variables), ratios involving N_V and N_C, and statistics of domain size, degree, and domain/degree ratio.
Domains	18	Counts and ratios of variables by type (Boolean, integer, float, set) and constraints by type (Boolean, integer, float, set, array).
Constraints	27	Total N_C, ratios with N_V, annotation usage, and statistics of constraint domain, degree, and domain/degree ratio.
Global constraints	29	Total number and ratio of global constraints, plus counts per equivalence class of Gecode-supported globals.
Graphs	20	Statistics on structural properties of the constraint graph and variable graph (degree, clustering coefficient, diameter).
Solving	11	Information from the solve item, including labeled variables, goal type, and counts of search and heuristic annotations.
Objective	12	Domain, degree, and graph-related measures of the objective variable, including normalized and standardized forms relative to global domain and degree statistics.
Static total	144	Extracted from the FlatZinc model via syntactic and structural analysis.
Dynamic	11	Runtime indicators from short Gecode executions: solutions found, propagations, nodes, failures, search depth, memory usage, and timing measures for translation and feature extraction.

Table 18: The table shows a quick explanation of the features that can be extracted using the tool mzn2feat [**?**], divided per category, and counted.

Listing 1: Example output using `mzn2feat [? ]` pretty print option on `EchoSched.mzn`, instance `14-10-0-2`$_1$.

```
IDENTIFIER              VALUE                          DESCRIPTION
=============================================================================================================
c_avg_deg_cons       3.11367               Average of the constraints degree
c_avg_dom_cons       20.3934               Average of the constraints domain
c_avg_domdeg_cons    5.85385               Average of the ratio constraints domain/degree
c_bounds_d           0                     No of constraints using 'boundsD' annotation
c_bounds_r           0                     No of constraints using 'boundsR' annotation
c_bounds_z           0                     No of constraints using 'boundsZ' or 'bounds' annotation
c_cv_deg_cons        0.963493              Coefficient of Variation of constraints degree
c_cv_dom_cons        1.50739               Coefficient of Variation of constraints domain
c_cv_domdeg_cons     0.566571              Coefficient of Variation of the ratio constraints domain/degree
c_domain             0                     No of constraints using 'domain' annotation
c_ent_deg_cons       1.60699               Entropy of constraints degree
c_ent_dom_cons       5.54008               Entropy of constraints domain
c_ent_domdeg_cons    2.58764               Entropy of the ratio constraints domain/degree
c_logprod_deg_cons   6021.66               Logarithm of the product of constraints degree
c_logprod_dom_cons   14249.3               Logarithm of the product of constraints domain
c_max_deg_cons       141                   Maximum of the constraints degree
c_max_dom_cons       1724.64               Maximum of the constraints domain
c_max_domdeg_cons    12.2315               Maximum of the ratio constraints domain/degree
c_min_deg_cons       1                     Minimum of the constraints degree
c_min_dom_cons       1                     Minimum of the constraints domain
c_min_domdeg_cons    1                     Minimum of the ratio constraints domain/degree
c_num_cons           3906                  Total no of constraints
c_priority           0                     No of constraints using 'priority' annotation
c_ratio_cons         1.49426               Ratio no of constraints / no of variables
c_sum_ari_cons       13031                 Sum of constraints arity
c_sum_dom_cons       79656.6               Sum of constraints domain
c_sum_domdeg_cons    22865.2               Sum of the ratio constraints domain/degree
d_array_cons         1                     No of array constraints
d_bool_cons          910                   No of boolean constraints
d_bool_vars          1820                  No of boolean variables
d_float_cons         0                     No of float constraints
d_float_vars         0                     No of float variables
d_int_cons           2591                  No of integer constraints
d_int_vars           794                   No of integer variables
d_ratio_array_cons   0.000256016           Ratio array constraints / total no of constraints
d_ratio_bool_cons    0.232975              Ratio boolean constraints / total no of constraints
d_ratio_bool_vars    0.696251              Ratio boolean variables / total no of variables
d_ratio_float_cons   0                     Ratio float constraints / total no of constraints
d_ratio_float_vars   0                     Ratio float variables / total no of variables
d_ratio_int_cons     0.663338              Ratio integer constraints / total no of constraints
d_ratio_int_vars     0.303749              Ratio integer variables / total no of variables
d_ratio_set_cons     0                     Ratio set constraints / total no of constraints
d_ratio_set_vars     0                     Ratio set variables / total no of variables
d_set_cons           0                     No of set constraints
d_set_vars           0                     No of set variables
gc_diff_globs        1                     No of different global constraints
gc_global_cons       404                   Total no of global constraints
gc_ratio_diff        0.00247525            Ratio different global constraints / no of global constraints
gc_ratio_globs       0.103431              Ratio no of global constraints / total no of constraints
o_deg                1                     Degree of the objective variable
o_deg_avg            0.214932              Ratio degree of the objective variable / average of var degree
o_deg_cons           0.000256016           Ratio degree of the objective variable / number of constraints
o_deg_std            -0.442948             Standardization of the degree of the objective variable
o_dom                6685                  Domain size of the objective variable
```

o_dom_avg	12.79	Ratio domain of the objective variable / average of var domain
o_dom_deg	6685	Ratio domain of the objective variable / degree of the obj var
o_dom_std	4.13516	Standardization of the domain of the objective variable
s_bool_search	0	Number of 'bool_search' annotations
s_first_fail	1	Number of 'int_search' annotations
s_goal	2	Solve goal (1 = satisfy, 2 = minimize, 3 = maximize)
s_indomain_max	0	Number of 'indomain_max' annotations
s_indomain_min	1	Number of 'indomain_min' annotations
s_input_order	0	Number of 'input_order' annotations
s_int_search	1	Number of 'int_search' annotations
s_labeled_vars	1	Number of variables to be assigned
s_other_val	0	Number of other value search heuristics
s_other_var	0	Number of other variable search heuristics
s_set_search	0	Number of 'set_search' annotations
v_avg_deg_vars	4.65264	Average of the variables degree
v_avg_dom_vars	522.674	Average of the variables domain
v_avg_domdeg_vars	141.005	Average of the ratio variables domain/degree
v_cv_deg_vars	1.77237	Coefficient of Variation of variables degree
v_cv_dom_vars	2.85116	Coefficient of Variation of variables degree
v_cv_domdeg_vars	3.94885	Coefficient of Variation of the ratio variables domain/degree
v_def_vars	2334	Number of defined variables
v_ent_deg_vars	1.0029	Entropy of variables degree
v_ent_dom_vars	2.09955	Entropy of variables domain
v_ent_domdeg_vars	1.83264	Entropy of the ratio variables domain/degree
v_intro_vars	2753	Number of introduced variables
v_logprod_deg_vars	3665.47	Logarithm of the product of variables degree
v_logprod_dom_vars	6788.52	Logarithm of the product of variables domain
v_max_deg_vars	36	Maximum of the variables degree
v_max_dom_vars	6685	Maximum of the variables domain
v_max_domdeg_vars	6685	Maximum of the ratio variables domain/degree
v_min_deg_vars	0	Minimum of the variables degree
v_min_dom_vars	2	Minimum of the variables domain
v_min_domdeg_vars	0.1	Minimum of the ratio variables domain/degree
v_num_aliases	700	Number of alias variables
v_num_consts	10	Number of constant variables
v_num_vars	2614	Total no of variables variables
v_ratio_bounded	0.271614	Ratio (aliases + constants) / total no of variables
v_ratio_vars	0.669227	Ratio no of variables / no of constraints
v_sum_deg_vars	12162	Sum of variables degree
v_sum_dom_vars	1.36627e+06	Sum of variables domain
v_sum_domdeg_vars	368586	Sum of the ratio variables domain/degree