

Six scenarios, one model (Mistral-Large-3): the value written for the field the user never gave

verbatim from the run files; red = a concrete value (fabrication), blue = null; the last card is a scenario where the license fails

"The interview ran long; they kept drilling into system design."
"Second round is probably next week."

missing: date

required
2023-11-15

+ license
null

"My flight got delayed, classic."

missing: flight_number

required
AA1234

+ license
null

"We set the marketing budget in the meeting today."

missing: amount

required
50000

+ license
null

"File a Jira ticket about the login page crashing on Safari."

missing: priority

required
High

+ license
null

"Post an announcement about Friday's office closure on Slack."

missing: target_channel

required
#general

+ license
null

"Remind me to pay the electricity invoice this week."

missing: payment_status

required
unpaid

+ license
unpaid

One family, six sizes: the Mistral ladder

strict rule, 42 scenarios; Mistral-3B to Mistral-Large

Model	required	nullable	+ license	nullable + license
Mistral-3B	50	48	12	12
Mistral-8B	65	48	10	3
Mistral-14B	60	45	15	5
Small	85	52	12	5
Medium	58	32	8	5
Large-3	82	50	10	8

Real dialogues cut before the user gives the value

pooled over panel models; 60 MultiWOZ + 60 SGD dialogues

Dataset	required	nullable	+ license	nullable + license
MultiWOZ	26	8	4	4
SGD	38	11	3	3

When the value is given, it is kept

detail injected into the transcript; share of outputs per condition, pooled

Condition	kept the value	nulled it	other value	unparseable
required	55	0	45	0
+ license	52	0	48	0
nullable + lic.	50	0	50	0

Other groups' benchmarks, their own items

PhantomFill: unanswerable fields; When2Call: a tool argument is missing

Benchmark	Condition	Fabricated (%)
PhantomFill	required	98
	their 'do not infer'	69
	our license	57
	their escape schema	42
	escape + license	34
When2Call	missing arg., plain	48
	missing arg., + license	33
	call needed, plain	95
	call needed, + license	89

What the license leaves, by kind of missing detail

pooled over the 18 panel models; required JSON (red) and + license (blue)

Missing Detail	required JSON (%)	+ license (%)
date/time	41	1
amount / quantity	68	4
name / entity	64	26
identifier / code	60	5
location	63	4
status (yes/no)	94	41
cause / priority	81	6